

Europäisches Patentamt

European Patent Office

Office européen des brevets



(11)

**EP 0 695 941 A2**

(12)

**EUROPEAN PATENT APPLICATION**(43) Date of publication:  
07.02.1996 Bulletin 1996/06(51) Int. Cl.<sup>5</sup>: **G01N 33/543**, C12Q 1/68,  
C07H 21/00

(21) Application number: 95303356.0

(22) Date of filing: 19.05.1995

(84) Designated Contracting States:  
CH DE FR GB IT LI NL

(30) Priority: 08.06.1994 US 255682

(71) Applicant: **AFFYMAX TECHNOLOGIES N.V.**  
Willemstad, Curaçao (AN)

(72) Inventors:

- Besemer, Donald M.  
Los Altos Hills, California 94022 (US)
- Goss, Virginia W.  
Santa Barbara, California 93109 (US)
- Winkler, James L.  
D313 Sunnyvale, California 94086 (US)

(74) Representative: **Mayes, Stuart David et al**  
London, EC4A 1PQ (GB)**(54) Method and apparatus for packaging a chip**

(57) A body 300 having a cavity 310 for mounting a substrate 120 fabricated with probe sequences at known locations according to the methods disclosed in United States Patent Number 5,143,854 and PCT WO 92/10092 or others, is provided. The cavity includes inlets 350 and

360 for introducing selected fluids into the cavity to contact the probes. Accordingly, a commercially feasible device for use in high throughput assay systems is provided

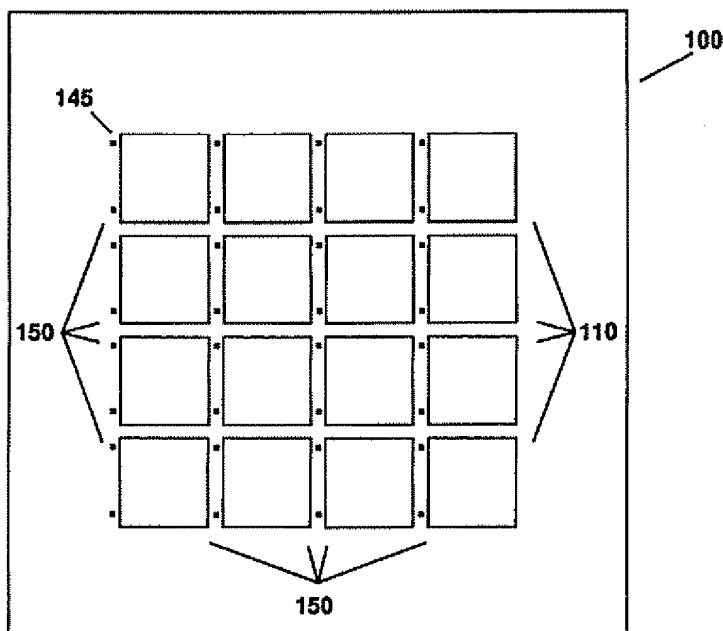


Figure 1a

EP 0 695 941 A2

1

EP 0 695 941 A2

2

## Description

## SUMMARY OF THE INVENTION

## BACKGROUND OF THE INVENTION

The present inventions relate to the fabrication and placement of materials at known locations on a substrate. In particular, one embodiment of the invention provides a method and associated apparatus for packaging a substrate having diverse sequences at known locations on its surface.

Techniques for forming sequences on a substrate are known. For example, the sequences may be formed according to the pioneering techniques disclosed in United States Patent Number 5,143,854 (Pirung et al.), PCT WO 92/10092, or United States Application Serial Number 08/249,188 (Attorney Docket Number 16528X-58), incorporated herein by reference for all purposes. The prepared substrates will have a wide range of applications. For example, the substrates may be used for understanding the structure-activity relationship between different materials or determining the sequence of an unknown material. The sequence of such unknown material may be determined by, for example, a process known as sequencing by hybridization. In one method of sequencing by hybridization, a sequences of diverse materials are formed at known locations on the surface of a substrate. A solution containing one or more targets to be sequenced is applied to the surface of the substrate. The targets will bind or hybridize with only complementary sequences on the substrate.

The locations at which hybridisation occurs can be detected with appropriate detection systems by labelling the targets with a fluorescent dye, radioactive isotope, enzyme, or other marker. Exemplary systems are described in United States Patent Number 5,143,854 (Pirung et al.) and United States Patent Application Serial Number 08/143,312, also incorporated herein by reference for all purposes. Information regarding target sequences can be extracted from the data obtained by such detection systems.

By combining various available technologies, such as photolithography and fabrication techniques, substantial progress has been made in the fabrication and placement of diverse materials on a substrate. For example, thousands of different sequences may be fabricated on a single substrate of about 1.28 cm<sup>2</sup> in only a small fraction of the time required by conventional methods. Such improvements make these substrates practical for use in various applications, such as biomedical research, clinical diagnostics, and other industrial markets, as well as the emerging field of genomics, which focuses on determining the relationship between genetic sequences and human physiology.

As commercialization of such substrates becomes widespread, an economically feasible and high-throughput device and method for packaging the substrates are desired.

Methods and devices for packaging a substrate having an array of probes fabricated on its surface are disclosed. In some embodiments, a body containing a cavity is provided. A substrate having an array of probes is attached to the cavity using, for example, an adhesive. The body includes inlets that allow fluids into and through the cavity. A seal is provided for each inlet to retain the fluid within the cavity. An opening is formed below the cavity to receive a temperature controller for controlling the temperature in the cavity. By forming a sealed thermostatically controlled chamber in which fluids can easily be introduced, a practical medium for sequencing by hybridization is provided.

In other embodiments, the body is formed by acoustically welding two pieces together. The concept of assembling the body from two pieces is advantageous. For example, the various features of the package (i.e., the channels, sealing means, and orientation means) are formed without requiring complex machining or designing. Thus, the packages are produced at a relatively low cost.

In connection with one aspect of the invention, a method for making the chip package is disclosed. In particular, the method comprises the steps of first forming a plurality of probe arrays on a substrate and separating the substrate into a plurality of chips. Typically, each chip contains at least one probe array. A chip is then mated to a package having a reaction chamber with fluid inlets. When mated, the probe array is in fluid communication with the reaction chamber.

A further understanding of the nature and advantages of the inventions herein may be realized by reference to the remaining portions of the specification and the attached drawings.

## BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1a illustrates a wafer fabricated with a plurality of probe arrays.

FIG. 1b illustrates a chip.

FIG. 2a illustrates a scribe and break device.

FIG. 2b illustrates the wafer mounted on a pick and place frame.

FIGS. 2c-2d illustrate the wafer, as displayed by the scribe and break device during alignment.

FIG. 3 illustrates a chip packaging device.

FIG. 4 illustrates the chip packaging device assembled from two components.

FIGS. 5a-5b illustrate the top and bottom view of a top casing of the chip packaging device.

FIG. 5c illustrates a different cavity orientation.

FIG. 6 illustrates a cross sectional view of the packaging device.

FIG. 7 illustrates the bottom view of a bottom casing of the chip packaging device.

FIGS. 8a-8b illustrate an acoustic welding system.

3

EP 0 695 941 A2

4

FIGS. 9a-9c illustrate the acoustic welding process used in assembling the chip packaging device.

FIG. 10 illustrates an adhesive dispensing system used in attaching the chip to the chip packaging device.

FIGS. 11-13 illustrate in greater detail the adhesive dispensing system of FIG. 10.

FIGS. 14a-14d illustrate the procedure for aligning the system of FIG. 10.

FIGS. 15a-15e illustrate images obtained during the alignment process of FIGS. 14a-14d.

FIGS. 16a-16b illustrate an alternative embodiment of a packaging device.

FIGS. 17a-17b illustrate another embodiment of a packaging device.

FIG. 18 illustrates an alternative embodiment for attaching the chip to the packaging device.

FIG. 19 illustrates another embodiment for attaching the chip to the packaging device.

FIGS. 20a-20b illustrate yet another embodiment for attaching the chip to the packaging device.

FIG. 21 illustrates an alternative embodiment for attaching the chip to the packaging device.

FIG. 22 illustrates another embodiment for attaching the chip to the packaging device.

FIG. 23 illustrates an alternative embodiment for sealing the cavity on the packaging device.

FIG. 24 illustrates another alternative embodiment for sealing the cavity on the packaging device.

FIG. 25 illustrates yet another embodiment for sealing the cavity on the packaging device.

FIGS. 26a-26b illustrate an alternative embodiment for sealing the cavity on the packaging device.

FIGS. 27a-27b illustrate an alternative embodiment for mounting the chip.

FIG. 28 illustrates an agitation system.

FIG. 29 illustrates an alternative embodiment of the agitation system.

FIG. 30 illustrates another embodiment of the agitation system.

## DESCRIPTION OF THE PREFERRED EMBODIMENT CONTENTS

### I. Definitions

### II. General

### III. Details of One Embodiment of Invention

#### a. Chip Package

#### b. Assembly of Chip Package

#### c. Chip Attachment

### IV. Details on Alternative Embodiments

#### a. Chip Package

#### b. Chip Attachment

#### c. Fluid Retention

#### d. Chip Orientation

#### e. Parallel Diagnostics

## V. Details of an Agitation System

### I. Definitions

The following terms are intended to have the following general meanings as they are used herein:

1. **Probe:** A probe is a surface-immobilized molecule that is recognized by a particular target and is sometimes referred to as a ligand. Examples of probes that can be investigated by this invention include, but are not restricted to, agonists and antagonists for cell membrane receptors, toxins and venoms, viral epitopes, hormones (e.g., opioid peptides, steroids, etc.), hormone receptors, peptides, enzymes, enzyme substrates, cofactors, drugs, lectins, sugars, oligonucleotides or nucleic acids, oligosaccharides, proteins, and monoclonal antibodies.

2. **Target:** A target is a molecule that has an affinity for a given probe and is sometimes referred to as a receptor. Targets may be naturally-occurring or manmade molecules. Also, they can be employed in their unaltered state or as aggregates with other species. Targets may be attached, covalently or non-covalently, to a binding member, either directly or via a specific binding substance. Examples of targets which can be employed by this invention include, but are not restricted to, antibodies, cell membrane receptors, monoclonal antibodies and antisera reactive with specific antigenic determinants (such as on viruses, cells or other materials), drugs, oligonucleotides or nucleic acids, peptides, cofactors, lectins, sugars, polysaccharides, cells, cellular membranes, and organelles. Targets are sometimes referred to in the art as anti-probes or anti-ligands. As the term "targets" is used herein, no difference in meaning is intended. A "Probe Target Pair" is formed when two macromolecules have combined through molecular recognition to form a complex.

### II. General

The present invention provides economical and efficient packaging devices for a substrate having an array of probes fabricated thereon. The probe arrays may be fabricated according to the pioneering techniques disclosed in United States Patent Number 5,143,854 (Pirung et al.), PCT WO 92/10092, or United States Application Serial Number \_\_\_\_\_ (Attorney Docket Number 16528X-58), already incorporated herein by reference for all purposes. According to one aspect of the techniques described therein, a plurality of probe arrays are immobilized at known locations on a large substrate or wafer.

Fig. 1a illustrates a wafer 100 on which numerous probe arrays 110 are fabricated. The wafer 100 may be composed of a wide range of material, either biological, nonbiological, organic, inorganic, or a combination of any of these, existing as particles, strands, precipitates,

5

EP 0 695 941 A2

6

gels, sheets, tubing, spheres, containers, capillaries, pads, slices, films, plates, slides, etc. The wafer may have any convenient shape, such as a disc, square, sphere, circle, etc. The wafer is preferably flat but may take on a variety of alternative surface configurations. For example, the wafer may contain raised or depressed regions on which a sample is located. The wafer and its surface preferably form a rigid support on which the sample can be formed. The wafer and its surface are also chosen to provide appropriate light-absorbing characteristics. For instance, the wafer may be a polymerized Langmuir Blodgett film, functionalized glass, Si, Ge, GaAs, GaP, SiO<sub>2</sub>, SiN<sub>4</sub>, modified silicon, or any one of a wide variety of gels or polymers such as (poly)tetrafluoroethylene, (poly)vinylidenedifluoride, polystyrene, polycarbonate, or combinations thereof. Other materials with which the wafer can be composed of will be readily apparent to those skilled in the art upon review of this disclosure. In a preferred embodiment, the wafer is flat glass or single-crystal silicon.

Surfaces on the solid wafer will usually, though not always, be composed of the same material as the wafer. Thus, the surface may be composed of any of a wide variety of materials, for example, polymers, plastics, resins, polysaccharides, silica or silica-based materials, carbon, metals, inorganic glasses, membranes, or any of the above-listed wafer materials.

Wafer 100 includes a plurality of marks 145 that are located in streets 150 (area adjacent to the probe arrays). Such marks may be used for aligning the masks during the probe fabrication process. In effect, the marks identify the location at which each array 110 is to be fabricated. The probe arrays may be formed in any geometric shape. In some embodiments, the shape of the array may be squared to minimize wasted wafer area. After the probe arrays have been fabricated, the wafer is separated into smaller units known as chips. The wafer, for example, may be about 5 x 5 inches on which 16 probe arrays, each occupying an area of about 12.8 cm<sup>2</sup>, are fabricated.

Fig. 1b illustrates a chip that has been separated from the wafer. As illustrated, chip 120 contains a probe array 110 and a plurality of alignment marks 145. The marks serve multiple functions, such as: 1) aligning the masks for fabricating the probe arrays, 2) aligning the scribe for separating the wafer into chips, and 3) aligning the chip to the package during the attachment process. In some embodiments, such chips may be of the type known as Very Large Scale Immobilized Polymer Synthesis (VLSIPS™) chips.

According to a specific embodiment, the chip contains an array of genetic probes, such as an array of diverse RNA or DNA probes. In some embodiments, the probe array will be designed to detect or study a genetic tendency, characteristic, or disease. For example, the probe array may be designed to detect or identify genetic diseases such as cystic fibrosis or certain cancers (such as P53 gene relevant to some cancers), as disclosed in

United States Patent Application Serial Number 08/143,312, already incorporated by reference.

According to one embodiment, the wafer is separated into a plurality of chips using a technique known as scribe and break. Fig. 2a illustrates a fully programmable computer controlled scribe and break device, which in some embodiments is a DX-III Scribe breaker manufactured by Dynatex International™. As shown, the device 200 includes a base 205 with a rotation stage 220 on which a wafer is mounted. The rotation stage includes a vacuum chuck for fixing the wafer thereon. A stepper motor, which is controlled by the system, rotates stage 220. Located above the stage is a head unit 230 that includes a camera 232 and cutter 231. Head unit 230 is mounted on a dual-axis frame. The camera generates an image of the wafer on video display 210. The video display 210 includes a cross hair alignment mark 215. The camera, which includes a zoom lens and a fiber optic light, allows a user to inspect the wafer on the video display 210. A control panel 240 is located on the base for operating device 200.

In operation, a user places a wafer 100 on a frame 210 as illustrated in Fig. 2b. The surface of frame 210 is composed of a flexible and sticky material. The tackiness of the frame prevents the chips from being dispersed and damaged during the breaking process. Frame 210 may be a pick and place frame or a hoop that is commonly associated with fabrication of semiconductors. Referring back to Fig. 2a, a user places the frame with the wafer on the rotation stage 220. In some embodiments, the frame is held on the rotation stage by vacuum pressure. The user then aligns the wafer by examining the image displayed on the video display 210.

According to one embodiment, wafer alignment is achieved in two steps. First, using the control panel 240, the user rotates stage 220. The stage is rotated until streets 150 are aligned with the cross hair 215 on the display, as illustrated in Fig. 2c. Next, the user moves the cutter until it is aligned at the center of one of the streets. This step is performed by aligning horizontal line 216 of the cross hair between alignment marks 145, as shown in Fig. 2d.

Once the cutter is aligned, the user instructs the device to scribe the wafer. In some embodiments, various options are available to the user, such as scribe angle, scribe pressure, and scribe depth. These parameters will vary depending on the composition and/or thickness of the wafer. Preferably, the parameters are set to scribe and break the wafer without causing any damage thereto or penetrating through the frame. The device repeatedly scribes the wafer until all the streets in one axis have been scribed, which in one embodiment is repeated 5 times (a 4x4 matrix of probe arrays). The user then rotates the stage 90° to scribe the perpendicular streets.

Once the wafer has been scribed, the user instructs the device to break or separate the wafer into chips. Referring back to Fig. 2a, the device 200 breaks the wafer by striking it beneath the scribe with an impulse

bar located under the rotation table 220. The shock from the impulse bar fractures the wafer along the scribe. Since most of the force is dissipated along the scribe, device 200 is able to produce high breaking forces without exerting significant forces on the wafer. Thus, the chips are separated without causing any damage to the wafer. Once separated, the chips are then packaged. Of course, other more conventional techniques, such as the sawing technique disclosed in U.S. Patent Number 4,016,855, incorporated herein by reference for all purposes, may be employed.

#### IV. Details of One Embodiment of the Invention

##### a. Chip Package

Fig. 3 illustrates a device for packaging the chips. Package 300 contains a cavity 310 on which a chip is mounted. The package includes inlets 350 and 360 which communicate with cavity 310. Fluids are circulated through the cavity via inlets 350 and 360. A septum, plug, or other seal may be employed to seal the fluids in the cavity. Alignment holes 330 and 335 may be provided for alignment purposes. In some embodiments, the package may include a non-flush edge 320. In some detection systems, the packages may be inserted into a holder similar to an audio cassette tape. The asymmetrical design of the package will assure correct package orientation when inserted into the holder.

Fig. 4 illustrates one embodiment of the package. As shown in Fig. 4, the chip package is manufactured by mating two substantially complementary casings 410 and 420 to form finished assembly 300. Preferably, casings 410 and 420 are made from injection molded plastic. Injection molding enables the casings to be formed inexpensively. Also, assembling the package from two parts simplifies the construction of various features, such as the internal channels for introducing fluids into the cavity. As a result, the packages may be manufactured at a relatively low cost.

Figs. 5a-5b show the top casing 410 in greater detail. Fig. 5a shows a top view and Fig. 5b shows a bottom view. Referring to Fig. 5a, top casing 410 includes an external planar surface 501 having a cavity 310 therein. In some embodiments, the surface area of casing 410 sufficiently accommodates the cavity. Preferably, the top casing is of sufficient size to accommodate identification labels or bar codes in addition to the cavity. In a specific embodiment, the top casing is about 1.5" wide, 2" long, and 0.2" high.

Cavity 310 is usually, though not always, located substantially at the center of surface 501. The cavity may have any conceivable size, shape, or orientation. Preferably, the cavity is slightly smaller than the surface area of the chip to be placed thereon and has a volume sufficient to perform hybridization. In one embodiment, the cavity may be about 0.58" wide, 0.58" long, and 0.2" deep.

Cavity 310 may include inlets 350 and 360. Selected fluids are introduced into and out of the cavity via the inlets. In some embodiments, the inlets are located at opposite ends of the cavity. This configuration improves fluid circulation and regulation of bubble formation in the cavity. The bubbles agitate the fluid, increasing the hybridization rate between the targets and complementary probe sequences. In one embodiment, the inlets are located at the top and bottom end of the cavity when the package is oriented vertically such as at the opposite corners of the cavity. Locating the inlet at the highest and lowest positions in the cavity facilitates the removal of bubbles from the cavity.

Fig. 5c illustrates an alternative embodiment in which cavity 310 is oriented such that the edges of the cavity 310 and the casing 410 are non-parallel. This configuration allows inlets 350 and 360 to be situated at the absolute highest and lowest locations in the cavity when the package is vertically oriented. As a result, bubbles or fluid droplets are prevented from being potentially trapped in the cavity.

Referring back to Fig. 5a, a depression 550 surrounds the cavity. In some embodiments, a ridge 560 may be provided at the edge of the depression so as to form a trough. The ridge serves to support the chip above the cavity. To attach the chip to the package, an adhesive may be deposited in the trough. This configuration promotes efficient use of chip surface area, thus increasing the number of chips yielded from a wafer.

Top casing 410 includes alignment holes 330 and 335. In some embodiments, holes 330 and 335 are different in size to ensure correct orientation of the package when mounted on an alignment table. Alternatively, the holes may have different shapes to achieve this objective. Optionally, the holes taper radially inward from surface 501 toward 502 to reduce the friction against alignment pins while still maintaining adequate contact to prevent slippage.

Referring to Fig. 5b, channels 551 and 561 are optionally formed on internal surface 502. Channels 551 and 561 communicate with inlets 350 and 360 respectively. A depression 590 is formed below cavity. According to some embodiments, the shape of depression 590 is symmetrical to the cavity with exception to corners 595 and 596, which accommodate the inlets. The depth of depression 590 may be, for example, about 0.7". As a result, the bottom wall of the cavity is about 0.05" thick. Depression 590 may receive a temperature controller to monitor and maintain the cavity at the desired temperature. By separating the temperature controller and cavity with a minimum amount of material, the temperature within the cavity may be controlled more efficiently and accurately. Alternatively, channels may be formed on surface 502 for circulating air or water to control the temperature within the cavity.

In some embodiments, certain portions 595 of internal surface 502 may be eliminated or cored without interfering with the structural integrity of the package when assembled. Coring the casing reduces the wall thick-

ness, causing less heat to be retained during the injection molding process; potential shrinkage or warpage of the casing is significantly reduced. Also, coring decreases the time required to cool the casing during the manufacturing process. Thus, manufacturing efficiency is improved.

In one embodiment, the top casing and bottom casing are mated together using a technique known as acoustic or ultrasonic welding. Accordingly, "energy directors" 510 are provided. Energy directors are raised ridges or points, preferably v-shaped, that are used in an acoustic welding process. The energy directors are strategically located, for example, to seal the channels without interfering with other features of the package and to provide an adequate bond between the two casings. Alternatively, the casings may be mated together by screws, glue, clips, or other mating techniques.

Figs. 6 shows a cross sectional view of the cavity 310 with chip 120 mounted thereon in detail. As shown, a depression 550 is formed around cavity 310. The depression includes a ridge 560 which supports chip 120. The ridge and the depression create a trough around cavity 310. In some embodiments, the trough is sufficiently large to receive an adhesive 630 for attaching the chip to the package. In one embodiment, the trough is about 0.08" wide and 0.06" deep. When mounted, the edge of the chip protrudes slightly beyond ridge 550, but without contacting side 625 of the depression. This configuration permits the adhesive to be dispensed onto the trough and provides adequate surface area for the adhesive to attach chip 120 to the package.

According to some embodiments, the back surface 130 of chip 120 is at least flush or below the plane formed by surface 501 of casing 410. As a result, chip 120 is shielded by surface 501 from potential damage. This configuration also allows the packages to be easily stored with minimal storage area since the surfaces are substantially flat.

Optionally, the bottom of the cavity includes a light absorptive material, such as a glass filter or carbon dye, to prevent impinging light from being scattered or reflected during imaging by detection systems. This feature improves the signal-to-noise ratio of such systems by significantly reducing the potential imaging of undesired reflected light.

Fig. 7 shows the internal surface of bottom casing 420 in greater detail. As shown, the bottom casing 420 is substantially planar and contains an opening 760 therein. Preferably, the casing 420 is slightly wider or slightly longer than the top casing. In one embodiment, casing 420 is about 1.6" wide, 2.0" long, and 0.1" deep, which creates a non-flush edge on the finish assembly. As previously mentioned, this design ensures that the package is correctly oriented when mounted onto the detection systems.

In some embodiments, opening 760 is spatially located at about the depression below the cavity. The opening also has substantially the same geometric configuration as the depression to allow the temperature

controller to contact as much of the bottom of the cavity as possible.

Internal surface 701 of casing 420 includes depressions 730 and 740. A port 731 is located in depression 730 and a port 741 is located in depression 740. Ports 731 and 741 communicate with channels on the top casing (350 and 360 in Fig. 5b) when the package is assembled. A seal 790, which may be a septum composed of rubber, teflon/rubber laminate, or other sealing material is provided for each depression. The septum may be of the type commonly used to seal and reseal vessels when a needle is inserted into the septum for addition/removal of fluids. The septums, when seated in the depressions, extend slightly above surface, which in some embodiments is about 0.01".

This design causes casings 410 and 420 to exert pressure on the septum, forming a seal between the ports and the channels. The seal is maintained even after fluid is injected into the cavity since the pressure immediately forces the septum to reseal itself after the needle or other fluid injecting means is removed from the port. Thus, an efficient and economical seal for retaining fluid in the cavity is provided.

Also, casing 420 includes the complementary half alignment holes 330 and 335, each tapering radially inward from the external surface. Further, certain areas 765 on internal surface 701 may be cored, as similar to the internal surface of the top casing.

#### b. Assembly of Chip Package

According to one embodiment, the top and bottom casing are attached by a technique known as ultrasonic or acoustic welding. Fig. 8a is a schematic diagram of acoustic welding system used for assembling the package. In some embodiments, the welding system 800 is a HS Dialog ultrasonic welder manufactured by Herrmann Ultrasonics Inc. System 800 includes a platform 850 mounted on base 810. Platform 850 accommodates the top and bottom casings during the assembling process.

An acoustic horn 860 is mounted on a frame above platform 850. The horn translates vertically (toward and away from platform 850) on the frame by air pressure. The horn is connected to a frequency generator 870, which in some embodiments is a 20 KHz generator manufactured by Herrmann Ultrasonics Inc. System 800 is controlled by a controller 880, which, for example, may be a Dialog 2012 manufactured by Herrmann Ultrasonics Inc. Controller 880 may be configured to accept commands from a digital computer system 890. Computer 890 may be any appropriately programmed digital computer of the type that is well known to those skilled in the art such as a Gateway 486DX operating at 33 MHz.

Fig. 8b illustrates platform 850 in greater detail. The platform 850 is substantially planar and includes alignment pins 851 and 852. Alignment pins 851 and 852 are used to align both the top and bottom casings during the welding process. In some embodiments, a pad 890,

11

EP 0 695 941 A2

12

which may be composed of silicone rubber or other energy absorbing material, is located on platform 850 to prevent damage to the package during assembly.

Fig. 9a illustrates the acoustic welding system in operation. As shown, bottom casing 420, having a septum 790 seated in each depression, is mounted onto platform table 850 and held in place by alignment pins. Top casing 410 is then aligned above the bottom casing with alignment pins. The system then commences the welding process by lowering horn 860 until it contacts the top surface of casing 410.

Fig. 9b illustrates the casing and horn in detail. As shown, the horn 860 presses against top casing 410, thereby forcing energy directors 510 to interface with bottom casing 420. The system then activates the frequency generator, causing the welding horn to vibrate.

Fig. 9c illustrates in detail the energy directors during the welding process. As shown in step 9001, welding horn 860 forces energy directors 510 against bottom casing 420. At step 9002, the system vibrates the welding horn, which in some embodiments is at 20 KHz. The energy generated by the horn melts the energy directors. Simultaneously, the horn translates downward against the package. At step 9003, the pressure exerted by the horn causes the energy directors to fuse with the bottom casing. At step 9004, the welding process is completed when the horn reaches its weld depth, for example, of about 0.01". Of course, the various welding parameters may be varied, according to the composition of the materials used, to achieve optimum results.

### c. Chip Attachment

According to some embodiments, an ultraviolet cured adhesive attaches the chip to the package. Fig. 10 schematically illustrates an adhesive dispensing system used in attaching the chip. The dispensing system 1000 includes an attachment table 1040 to accommodate the package during the attachment process. A chip alignment table 1050 for aligning the chip is located adjacent to attachment table 1040. A head unit 1030 for dispensing the adhesive is located above tables 1040 and 1050. The head unit 1030 also includes a camera that generates an output to video display 1070. Video display 1070, in some embodiments, includes a cross hair alignment mark 1071. The head unit is mounted on a dual-axis (x-y) frame for positioning during alignment and attachment of the chip. The operation of the dispensing system is controlled by a computer 1060, which in some embodiments may be Gateway 486DX operating at 33 MHz.

Fig. 11 illustrates the attachment table in greater detail. The attachment table 1040 has a substantially flat platform 1110 supported by a plurality of legs 1105. Alignment pins 1115 and 1116, which secure the package during the attachment process, are located on the surface of platform 1110.

Optionally, a needle 1120 is provided. Needle 1120 includes a channel 1121 and is connected to a vacuum pump. In operation, the needle is inserted into one of the

ports of the package in order to generate a vacuum in the cavity. The vacuum pressure secures the chip to the package during the attachment process.

Fig. 12a shows table 1050 in greater detail. Table 1050 includes a substantially flat platform 1210 having a depression 1240 for holding a chip. In some embodiments, a port 1241 is provided in depression 1240. Port 1241 is connected to a vacuum pump which creates a vacuum in the depression for immobilizing the chip therein. Platform 1210 is mounted on a combination linear rotary stage 1246, which in some embodiments may be a model 26LR manufactured by DARDAL, and a single axis translation stage 1245, which may be a model CR2226HSE2 manufactured by DARDAL.

Fig. 12b illustrates depression 1240 in greater detail. As shown, a ledge 1241 surrounds the depression 1240. Ledge 1241 supports the chip when it is placed above depression 1240. Since the chips are placed over the depression with the probes facing the table, this design protects the probes from being potentially damaged during alignment.

Fig. 13 illustrates the head unit 1030 in greater detail. As shown, the head unit 1030 includes a camera assembly 1320 that generates an output to a video display. A light 1360 is provided to enable the camera to focus and image an object of interest. The head unit also includes an ultraviolet light 1350 for curing the adhesive, a vacuum pickup 1330 for moving chip during the attachment process, and an adhesive dispenser 1340.

In operation, a chip package is placed onto table 1040. As previously described, the alignment pins on the table immobilize the package. The user begins the chip attachment process by calibrating the head unit. This may be done by moving the camera above the package and aligning it with a mark on the package, as shown in Fig. 14a. For convenience, one of the alignment pins may be used as an alignment mark. Fig. 14b illustrates a typical image 1440 generated by the camera during this step. As shown, the head unit is not aligned with pin 1480. To align the head unit, the user translates it in both the x and y direction until pin 1480 is located at the intersection 1477 of the cross hair on the video display, as illustrated in Fig. 14c.

Next, the chip is inserted into the depression on the chip alignment table. Fig. 14c is a flow chart indicating the steps for aligning the chip. At step 1410, the system positions the camera (head unit) above one of the chip's alignment marks. The camera images the alignment mark on the video display. At this point, the mark is normally misaligned (i.e., the mark is not located at the intersection of the cross hair alignment mark). At step 1420, the user adjusts the chip alignment table in both the x and y direction until the mark is substantially located at the intersection of the cross hair. Since no rotational adjustments were made, the mark may be misaligned angularly.

At step 1430, the user instructs the system to move the camera above a second alignment mark, which usually is at an opposite corner of the chip. Again, an image

13

EP 0 695 941 A2

14

of the alignment mark is displayed. At this stage, the alignment mark is probably misaligned in the x, y, and angular directions. At step 1440, the user adjusts the rotational stage, x-stage, and y-stage, if necessary, to align the mark with the cross hair on the video display. In instances where the rotational stage has been rotated, the first alignment mark will become slightly misaligned. To compensate for this shift, the user repeats the alignment process beginning at step 1450 until both marks are aligned. Of course, image processing techniques may be applied for automated head unit and chip alignment.

Fig. 15a is an example of an image displayed by the video screen during step 1410. As shown, the first alignment mark (lower left corner of the chip) is not aligned with the cross hair marking. Fig. 15b exemplifies an image of the first alignment mark after adjustments were made by the user. Fig. 15c illustrates a typical image displayed by video screen during step 1430. As illustrated, the second alignment mark (upper right corner of the chip) is misaligned in the x, y, and angular directions. Fig. 15d illustrates an image of the second mark following initial adjustments by the user at step 1440. Fig. 15e illustrates the orientation of the second alignment mark after the chip has been aligned.

Once the chip is aligned, the vacuum holding the chip on the attachment table is released. Thereafter, the pickup on the head unit removes the chip from the table and aligns it on the cavity of the package. In some embodiments, the chip is mated to the pickup by a vacuum.

Optionally, the user may check to ensure that the chip is correctly aligned on the cavity by examining the chip's alignment marks with the camera. If the chip is out of position, the chip is removed and realigned on the alignment table. If the chip is correctly positioned, the system deposits an adhesive by moving the dispenser along the trough surrounding the cavity. In some embodiments, the vacuum is released before depositing the adhesive in the trough. This step is merely precautionary and implemented to ensure that the vacuum does not cause any adhesive to seep into the cavity. Once the adhesive is deposited, the system reexamines the chip to determine if the adhesive had moved the chip out of position. If the chip is still aligned, the head unit locates the ultraviolet light above the adhesive and cures it for a time sufficient to harden the adhesive, which in one embodiment is about 10 seconds. Otherwise, the chip is realigned.

Upon completion, the chip package will have a variety of uses. For example, the chip package will be useful in sequencing genetic material by hybridization. In sequencing by hybridization, the chip package is mounted on a hybridization station where it is connected to a fluid delivery system. Such system is connected to the package by inserting needles into the ports and puncturing the septums therein. In this manner, various fluids are introduced into the cavity for contacting the probes during the hybridization process.

Usually, hybridization is performed by first exposing the sample with a prehybridization solution. Next, the sample is incubated under binding conditions with a solution containing targets for a suitable binding period. Binding conditions will vary depending on the application and are selected in accordance with the general binding methods known including those referred to in: Maniatis et al., Molecular Cloning: A Laboratory Manual (1989), 2nd Ed., Cold Spring Harbor, N.Y. and Berger and Kimmel, Methods in Enzymology, Volume 152, Guide to Molecular Cloning Techniques (1987), Academic Press, Inc., San Diego, CA.; Young and Davis (1983) Proc. Natl. Acad. Sci. (U.S.A.) 80: 1194, which are incorporated herein by reference. In some embodiments, the solution may contain about 1 molar of salt and about 1 to 50 nanomolar of targets. Optionally, the fluid delivery system includes an agitator to improve mixing in the cavity, which shortens the incubation period. Finally, the sample is washed with a buffer, which may be 6X SSPE buffer, to remove the unbound targets. In some embodiments, the cavity is filled with the buffer after washing the sample.

Thereafter, the package may be aligned on a detection or imaging system, such as those disclosed in United States Patent Number 5,143,854 (Pirrung et al.) or United States Patent Application Serial Number 08/495,889 (Attorney Docket Number 11509-117), already incorporated herein by reference for all purposes. Such detection systems may take advantage of the package's asymmetry (i.e., non-flush edge) by employing a holder to match the shape of the package specifically. Thus, the package is assured of being properly oriented and aligned for scanning. The imaging systems are capable of qualitatively analyzing the reaction between the probes and targets. Based on this analysis, sequence information of the targets is extracted.

### III. Details on Alternative Embodiments

#### a. Chip Package Orientation

Figs. 16a-16b illustrate an alternative embodiment of the package. Fig. 16a shows a top view and Fig. 16b shows a bottom view. As shown in Fig. 16a, a cavity 1620 is located on a top surface 1610 of the package body 1600. The body includes alignment holes 1621 and 1622 that are used, for example, in mating the chip to the package. Optionally, a plurality of ridges 1690 is located at end 1660 of the body. The friction created by ridges 1690 allows the package to be handled easily without slipping.

The body also includes two substantially parallel edges 1630 and 1640. As shown, edge 1640 is narrowed at end 1665 to create an uneven edge 1645. The asymmetrical design of the body facilitates correct orientation when mounted onto detection systems. For example, detection systems may contain a holder, similar to that of an audio cassette tape, in which end 1665 is inserted.

15

EP 0 695 941 A2

16

Referring to Fig. 16b, ports 1670 and 1671 communicate with cavity 1620. A seal is provided for each port to retain fluids in the cavity. Similar to the top surface, the bottom surface may optionally include a plurality of ridges 1690 at end 1660.

Figs. 17a-17b illustrate an alternative embodiment of the package. Fig. 17a shows a top view and Fig. 17b shows a bottom view. Referring to Fig. 17a, a cavity 1720 is located on a top surface 1710 of the package body 1700. The body may be formed in the shape of a disk with two substantially parallel edges 1730 and 1740. Alignment holes 1721 and 1722, which may be different in size or shape, are located on the body. In some embodiments, the package is inserted like an audio cassette tape into detection systems in a direction parallel to edges 1730 and 1740. Edges 1730 and 1740 and alignment holes prevent the package from being inserted incorrectly into the detection systems.

As shown in Fig. 17b, ports 1730 and 1740 are located on the bottom surface 1715 of the package. Ports 1730 and 1740 communicate with cavity 1720 and each include a seal 1780 for sealing fluids in the cavity.

#### b. Chip Attachment

Fig. 18 illustrates an alternative embodiment for attaching the chip to the package. As shown, two concentric ledges 1810 and 1820 surround the perimeter of cavity 310. Ledge 1820 supports the chip 120 when mounted above cavity 310. Ledge 1810, which extends beyond chip 120, receives an adhesive 1860 such as ultraviolet cured silicone, cement, or other adhesive for attaching the chip thereto.

Fig. 19 illustrates another embodiment for attaching the chip to the package. According to this embodiment, a ledge 1910 is formed around cavity 310. Preferably, the ledge is sufficiently large to accommodate an adhesive 1920 such as an adhesive film, adhesive layer, tape, or any other adhesive layer. Chip 120 attaches to the package when it contacts the adhesive film.

Fig. 20a illustrates yet another embodiment for attaching a chip to the package. As shown, a clamp 2010, such as a frame having a plurality of fingers 2015, attaches the chip to the package. Fig. 20b illustrates a cross sectional view. A ridge 2020 on surface 501 surrounds cavity 310. The ridge includes a ledge 2025 upon which chip 120 rests. Optionally, a gasket or a seal 2070 is located between the ledge and chip to ensure a tight seal around cavity 310. Clamp 2010 is attached to side 2040 of ridge 2020 and surface 501. In some embodiments, clamp 2010 is acoustically welded to the body. Accordingly, clamp 2010 includes energy directors 2050 located at its bottom. Alternatively, screws, clips, adhesives, or other attachment techniques may be used to mate clamp 2010 to the package. When mated, fingers 2015 secure chip 120 to the package.

Fig. 21 illustrates an alternative embodiment for attaching the chip to the package. A ridge 2110, having a notch 2115 at or near the top of ridge 2110, encom-

passes the cavity 310. Chip 120 is wedged and held into position by notch 2115. Thereafter, a process known as heat staking is used to mount the chip. Heat staking includes applying heat and force at side 2111 of ridge, thus forcing ridge tightly against or around chip 120.

Fig. 22 shows another embodiment of attaching a chip onto a package. As shown, a channel 2250 surrounds cavity 310. A notch 2240 for receiving the chip 120 is formed along or near the top of the cavity 310. In some embodiments, a gasket or seal 2270 is placed at the bottom of the notch to ensure a tight seal when the chip is attached. Once the chip is located at the notch, a V-shaped wedge 2260 is inserted into channel 2250. The wedge forces the body to press against chip's edges and seal 2260, thus mating the chip to the package. This process is known as compression sealing. Other techniques such as insert molding, wave soldering, surface diffusion, laser welding, shrink wrap, o-ring seal, surface etching, or heat staking from the top may also be employed.

#### c. Fluid Retention

Fig. 23 shows an alternative embodiment of package that employs check valves to seal the inlets. As shown, depressions 2305 and 2315 communicate with cavity 310 through inlets 350 and 360. Check valves 2310 and 2320, which in some embodiments may be duck-billed check valves, are seated in depressions 2305 and 2315. To introduce a fluid into the cavity, a needle is inserted into the check valve. When the needle is removed, the check valve reseals itself to prevent leakage of the fluid.

Fig. 24 illustrates another package that uses reusable tape for sealing the cavity 310. As shown, a tape 2400 is located above inlets 350 and 360. Preferably, end 2430 of tape is permanently fixed to surface 2480 while end 2410 remains unattached. The mid section 2420 of the tape is comprised of non-permanent adhesive. This design allows inlets to be conveniently sealed or unsealed without completely separating the tape from the package.

Fig. 25 illustrates yet another embodiment of the package that uses plugs to retain fluids within the cavity. As shown, depressions 2520 and 2530 communicate with cavity 310 via inlets 350 and 360. A plug 2510, which in some embodiment may be composed of rubber or other sealing material, is mated to each of the depressions. Plugs 2510 are easily inserted or removed for sealing and unsealing the cavity during the hybridization process.

Fig. 26a illustrates a package utilizing sliding seals for retaining fluids within the cavity. The seals are positioned in slots 2610 that are located above the inlets. The slots act as runners for guiding the seals to and from the inlets. Fig. 26b illustrates the seal in greater detail. Seal 2640, which may be composed of rubber, teflon rubber, or other sealing material, is mated to each slot 2610. The seal includes a handle 2650 which extends through the

17

EP 0 695 941 A2

18

slot. Optionally, the bottom of the seal includes an annular protrusion 2645 to ensure mating with inlet 350. The inlet is sealed or unsealed by positioning the seal appropriately along the slot. Alternatively, spring loaded balls, rotary ball valves, plug valves, or other fluid retention techniques may be employed.

#### d. Chip Orientation

Figs. 27a-27b illustrate an alternative embodiment of the package. Fig. 27a illustrates a top view and Fig. 27b shows a cross sectional view. As shown, package 2700 includes a cavity 2710 on a surface 2705. A chip 2790 having an array of probes 2795 on surface 2791 is mated to the bottom of cavity 2710 with an adhesive 2741. The adhesive, for example, may be silicone, adhesive tape, or other adhesive. Alternatively, clips or other mounting techniques may be employed. Optionally, the bottom of the cavity may include a depression in which a chip is seated.

This configuration provides several advantages such as: 1) permitting the use of any type of substrate (i.e., non-transparent or non-translucent), 2) yielding more chips per wafer since the chip does not require an edge for mounting, and 3) allowing chips of various sizes or multiple chips to be mated to the package.

A cover 2770 is mated to the package for sealing the cavity. Preferably, cover 2770 is composed of a transparent or translucent material such as glass, acrylic, or other material that is penetrable by light. Cover 2770 may be mated to surface 2705 with an adhesive 2772, which in some embodiments may be silicone, adhesive film, or other adhesive. Optionally, a depression may be formed around the cavity such that surface 2271 of the cover is at least flush with surface 2705. Alternatively, the cover may be mated to surface 2705 according to any of the chip attachment techniques described herein.

Inlets 2750 and 2751 are provided and communicate with cavity 2710. Selected fluids are circulated through the cavity via inlets 2750 and 2751. To seal the fluids in the cavity, a septum, plug, or other seal may be employed. In alternative embodiments, any of the fluid retention techniques described herein may be utilized.

#### e. Parallel Hybridization and Diagnostics

In an alternative embodiment, the body is configured with a plurality of cavities. The cavities, for example, may be in a 96-well micro-titre format. In some embodiments, a chip is mounted individually to each cavity according to the methods described above. Alternatively, the probe arrays may be formed on the wafer in a format matching that of the cavities. Accordingly, separating the wafer is not necessary before attaching the probe arrays to the package. This format provides significant increased throughput by enabling parallel testing of a plurality of samples.

#### V. Details of an Agitation System

Fig. 28 illustrates an agitation system in detail. As shown, the agitation system 2800 includes two liquid containers 2810 and 2820, which in the some embodiments are about 10 milliliters each. Container 2810 communicates with port 350 via tube 2850 and container 2820 communicates with port 360 via tube 2860. An inlet port 2812 and a vent port 2811 are located at or near the top of container 2810. Container 2820 also includes an inlet port 2822 and a vent 2821 at or near its top. Port 2812 of container 2810 and port 2822 of container 2820 are both connected to a valve assembly 2828 via valves 2840 and 2841. An agitator 2801, which may be a nitrogen gas ( $N_2$ ) or other gas, is connected to valve assembly 2828 by fitting 2851. Valves 2840 and 2841 regulate the flow of  $N_2$  into their respective containers. In some embodiments, additional containers (not shown) may be provided, similar to container 2810, for introducing a buffer and/or other fluid into the cavity.

In operation, a fluid is placed into container 2810. The fluid, for example, may contain targets that are to be hybridized with probes on the chip. Container 2810 is sealed by closing port 2811 while container 2820 is vented by opening port 2821. Next,  $N_2$  is injected into container 2810, forcing the fluid through tube 2850, cavity 310, and finally into container 2820. The bubbles formed by the  $N_2$  agitate the fluid as it circulates through the system. When the amount of fluid in container 2810 nears empty, the system reverses the flow of the fluid by closing valve 2840 and port 2821 and opening valve 2841 and port 2811. This cycle is repeated until the reaction between the probes and targets is completed.

In some applications, foaming may occur when  $N_2$  interacts with the fluid. Foaming potentially inhibits the flow of the fluid through the system. To alleviate this problem, a detergent such as CTAB may be added to the fluid. In one embodiment, the amount of CTAB added is about 1 millimolar. Additionally, the CTAB affects the probes and targets positively by increasing the rate at which they bind, thus decreasing the reaction time required.

The system described in Fig. 28 may be operated in an alternative manner. According to this technique, back pressure formed in the second container is used to reverse the flow of the solution. In operation, the fluid is placed in container 2810 and both ports 2811 and 2821 are closed. As  $N_2$  is injected into container 2810, the fluid is forced through tube 2850, cavity 310, and finally into container 2820. Because the vent port in container 2820 is closed, the pressure therein begins to build as the volume of fluid and  $N_2$  increases. When the amount of fluid in container 2810 nears empty, the flow of  $N_2$  into container 2810 is terminated by closing valve 2840. Next, the circulatory system is vented by opening port 2811 of container 2810. As a result, the pressure in container 2820 forces the solution back through the system toward container 2810. In one embodiment, the system is injected with  $N_2$  for about 3 seconds and vented for about

3 seconds. This cycle is repeated until hybridization between the probes and targets is completed.

Fig. 29 illustrates an alternative embodiment of the agitation system. System 2900 includes a vortexer 2910 on which the chip package 300 is mounted. A container 2930 for holding the fluid communicates with inlet 350 via tube 2950. A valve 2935 may be provided to control the flow of solution into the cavity. In some embodiments, circulator 2901, which may be a N<sub>2</sub> source or other gas source, is connected to container 2930. Alternatively, a pump or other fluid transfer device may be employed. The flow of N<sub>2</sub> into container 2930 is regulated by a valve 2936. Circulator 2901 is also connected to inlet tube 2950 via a valve 2902.

A waste container 2920 communicates with port 360 via outlet tube 2955. In one embodiment, a liquid sensor 2940 may be provided for sensing the presence of liquid in outlet tube 2955. Access to the waste container may be controlled by a valve 2921. Optionally, additional containers (not shown), similar to container 2930, may be employed for introducing a buffer or other fluid into the cavity.

The system is initialized by closing all valves and filling container 2930 with, for example, a fluid containing targets. Next, valves 2936, 2935, and 2955 are opened. This allows N<sub>2</sub> to enter container 2930 which forces the fluid to flow through tube 2950 and into the cavity. When the cavity is filled, valves 2935, 2936, and 2955 are closed to seal the fluid in the cavity. Next, the vortexer is activated to vibrate the chip package, similar to a paint mixer. In some embodiments, the vortexer may vibrate the package at about 3000 cycles per minutes. The motion mixes the targets in the fluid, shortening the incubation period. In some embodiments, the vortexer rotates the chip package until hybridization is completed. Upon completion, valve 2902 and 2955 are opened to allow N<sub>2</sub> into the cavity. The N<sub>2</sub> empties the fluid into waste container 2920. Subsequently, the cavity may be filled with a buffer or other fluid.

Fig. 30 illustrates an alternative embodiment in which the agitation system is partially integrated into the chip package. As shown, chip package 300 includes a cavity 310 on which the chip is mounted. Cavity 310 is provided with inlets 360 and 350. The package also includes chambers 3010 and 3020. A port 3021 is provided in chamber 3010 and is connected to inlet 360 by a channel 3025.

Chamber 3010 is equipped with ports 3011 and 3012. Port 3012 communicates with inlet 350 through a channel 3015. Channel 3015 is provided with a waste port 3016 that communicates with a fluid disposal system 3500 via a tube 3501. A valve 3502 regulates the flow of fluids into the disposal system. In some embodiments, the disposal system includes a waste container 3510 and fluid recovery container 3520 which are connected to tube 3501. A valve 3530 is provided to direct the flow of fluids into either the waste container or recovery container.

Port 3011 is coupled to a fluid delivery system 3600 through a tube 3601. Fluids flowing into chamber 3010 from the fluid delivery system are regulated by a valve 3602. The fluid delivery system includes fluid containers 3610 and 3620 that are interconnected with a tube 3690. Container 3610, which may hold a fluid containing targets, includes ports 3616 and 3615. Port 3616 is connected to tube 3690. A valve 3612 controls the flow of the fluid out of container 3610. A circulator 3605, which may be a N<sub>2</sub> source, is connected to port 3615 of container 3610. Alternatively, any type of gas, pump or other fluid transfer device may be employed. The flow of N<sub>2</sub> into container 3610 is controlled by a valve 3618. A valve 3619 may also be provided to vent container 3610.

Container 3620, which may hold a buffer, is provided with ports 3625 and 3626. Circulator 3605 is connected to port 3625. A valve 3621 is provided to control the flow of N<sub>2</sub> into container 3620. Port 3626 is connected to tube 3690 via a valve 3622. Valve 3622 regulates the flow of the buffer out of container 3620. Optionally, additional containers (not shown), similar to container 3620, may be configured for introducing other fluids into the cavity. A valve 3690 connects circulator 3605 to tube 3690 for controlling the flow of N<sub>2</sub> directly into the package. A valve 3652 is provided for venting the fluid delivery system.

In the initial operating state, all valves are shut. To start the hybridization process, a fluid containing targets is introduced into chamber 301 by opening valves 3602, 3612 and 3618. This injects N<sub>2</sub> into container 3610 which forces the fluid to flow through 3601 and into chamber 3010. When chamber 3010 is filled, valves 3612 and 3618 are closed. Next, valve 3642 is opened, allowing N<sub>2</sub> to flow directly into chamber 3010. The N<sub>2</sub> agitates and circulates the fluid into cavity 310 and out to chamber 3020. As the volume of fluid and N<sub>2</sub> in chamber 3020 increase, likewise does the pressure therein. When chamber 3020 approaches its capacity, valve 3642 is closed to stop the fluid flow. Thereafter, the system is vented by opening valve 3652. Venting the system allows the back pressure in chamber 3020 to reverse the flow of fluids back into chamber 3010. When chamber 3010 is filled, valve 3652 is closed and valve 3642 is opened to reverse the fluid flow. This cycle is repeated until hybridization is completed.

When hybridization is completed, the system may be drained. This procedure depends on which chamber the fluid is located in. If the fluid is located in chamber 3020, then valve 3502 is opened, while valve 3530 is positioned to direct the fluid into the appropriate container (recovery or waste). The pressure in chamber 3020 forces the fluid through port 3016, tube 3501, and into the disposal system. If the fluid is in chamber 3010, then valve 3502 and 3642 are opened. As a result, N<sub>2</sub> forces the fluid in chamber 3010 through port 3501 and into the disposal system.

Once the system is emptied, all valves are closed. A buffer or other fluid may be introduced into the cavity. For example, the cavity may be filled with a buffer by

21

EP 0 695 941 A2

22

opening valves 3601, 3621, and 3622. This injects N<sub>2</sub> into container 3620 which forces the buffer therein to flow through the system until it fills cavity 310. In the alternative, ultrasonic radiation, heat, magnetic beads, or other agitation techniques may be employed.

The present inventions provide commercially feasible devices for packaging a probe chip. It is to be understood that the above description is intended to be illustrative and not restrictive. Many embodiments will be apparent to those skilled in the art upon reviewing the above description. Merely as an example, the package may be molded or machined from a single piece of material instead of two. Also, other asymmetrical designs may be employed to orient the package onto the detection systems.

The scope of the invention should, therefore, be determined not with reference to the above description, but instead should be determined with reference to the appended claims along with their full scope of equivalents.

### Claims

1. A method of making probe chips comprising the steps of:
  - forming a plurality of probe arrays on a substrate;
  - separating said substrate into a plurality of chips, each of said chips comprising at least one probe array thereon; and
  - mating at least one of said chips to a package, said package comprising a reaction chamber, said reaction chamber comprising inlets for flowing fluid therein, said at least one probe array in fluid communication with said reaction chamber.
2. The method as recited in claim 1 wherein said package is made by the steps of:
  - injection molding first and second halves of said package; and
  - mating said first and second halves together.
3. The method as recited in claim 2 wherein said halves are injection molded plastic.
4. The method as recited in claim 3 wherein one of said halves comprises flow channels therein, said flow channels in communication with said inlets.
5. The method as recited in claim 4 further comprising the step of applying a reenterable seal to flow channels in said package.
6. The method as recited in claim 1 wherein said substrate comprises alignment marks for forming said probe arrays thereon in a desired position, and wherein said alignment marks are used to identify locations for said separating of said substrate into chips.
7. The method as recited in claim 1 wherein said substrate comprises alignment marks for forming said probe arrays thereon in a desired position, and wherein said step of mating said chips to said package uses said alignment marks for positioning said chips at a desired location on said package.
8. The method as recited in claim 1 wherein said package comprises an alignment structure thereon, wherein said step of mating said chip to said package uses said alignment structures to position said package at a desired position.
9. The method as recited in claim 1 wherein said package comprises an alignment structure thereon, and further comprising the step of identifying the location of at least one target on said probe array in a scanner, wherein said package is placed at a desired location in said scanner using said alignment structure.
10. The method as recited in claim 9 wherein said alignment structure is a plurality of holes in said package.
11. The method as recited in claim 1 wherein said step of forming a plurality of probe arrays comprises the steps of:
  - selectively exposing said substrate to light;
  - coupling selected monomers to said substrate where said substrate has been exposed to light.
12. The method as recited in claim 1 wherein said step of separating comprises the steps of:
  - scribing said substrate in desired locations;
  - breaking said substrate along said scribe lines.
13. The method as recited in claim 1 wherein said step of forming a plurality of probe arrays on said substrate is a step of forming a plurality of oligonucleotide probe arrays on said substrate.
14. The method as recited in claim 13 further comprising the steps of flowing labeled oligonucleotide target molecules through said reaction chamber and identifying where said target molecules have bound to said substrate.
15. The method as recited in claim 14 wherein said package comprises a temperature probe and further comprising the step of monitoring and adjusting a temperature in said reaction chamber.
16. The method as recited in claim 1 wherein said package is formed by the steps of:
  - forming first and second package portions;
  - and

23

EP 0 695 941 A2

24

- acoustically welding said first and second package portions together.
17. The method as recited in claim 1 wherein said step of mating said chips to packages comprises the step of binding said chips to said package with an adhesive. 5
18. The method as recited in claim 17 wherein said packages comprise a recessed region thereon, whereby said chips do not extend above a surface of said packages. 10
19. The method as recited in claim 1 further comprising the step of flowing target molecules through said reaction chamber. 15
20. The method as recited in claim 19 wherein said step of flowing is a step of flowing material from an inlet along a first diagonal of a generally rectangular reaction chamber to an outlet along said diagonal of said generally rectangular reaction chamber. 20
21. The method as recited in claim 19 wherein said step of flowing target molecules through said reaction chamber comprises the steps of piercing an inlet septum and flowing said target molecules in a fluid through said reaction chamber. 25
22. An apparatus for packaging a substrate, said apparatus comprising: 30  
     a substrate having a first surface and a second surface, said first surface comprising a probe array;  
     a body having a mounting surface with a fluid cavity, said second surface attached to said cavity; and 35  
     a cover attached to said mounting surface for sealing said cavity. 40
23. The apparatus of claim 22 wherein said cavity comprises an inlet port and an outlet port, said inlet and outlet ports permitting fluids to circulate into and through said cavity 45
24. The apparatus of claim 23 wherein said inlet and outlet ports comprise a reenterable seal.
25. The apparatus of claim 22 wherein said probe array comprises an array of oligonucleotide probes 50
26. The apparatus of claim 22 wherein said body comprises a depression for receiving a temperature controller for controlling and maintaining the reaction temperature in said cavity. 55
27. The apparatus of claim 22 wherein said body comprises a first half mated to a second half.
28. The apparatus of claim 27 wherein said first and second halves are injection molded plastic.
29. The apparatus of claim 28 wherein one of said halves comprises a cavity, said cavity having an inlet port and an outlet port.
30. The apparatus of claim 29 wherein one of said halves comprises flow channels, said flow channels in communication with said inlet and outlet ports when said first and second halves are mated together.
31. The apparatus of claim 30 wherein said flow channels comprise a reenterable seal for sealing fluid in said cavity.
32. The apparatus of claim 22 wherein said body further comprises an alignment structure, said alignment structure used for maintaining said body in a desired position when mounted on detection system.
33. The apparatus of claim 32 wherein said alignment structure comprises a plurality of holes.
34. The apparatus of claim 32 wherein said alignment structure comprises a non-flush edge.
35. A method of evaluating probe chips comprising the steps of:  
     forming a plurality of probe arrays on a substrate;  
     separating said substrate into a plurality of chips, each of said chips comprising at least one probe array thereon;  
     mating at least one of said chips to a package, said package comprising a reaction chamber, said reaction chamber in fluid communication with an inlet and outlet, said at least one probe array in fluid communication with said reaction chamber; and  
     flowing labeled target molecules into said reaction chamber, said labeled target molecules reacting with said at least one probe array.
36. The method as recited in claim 35 wherein said step of forming a plurality of probe arrays is a step of forming a plurality of oligonucleotide probe arrays.
37. The method as recited in claim 35 wherein said inlet and said outlet comprise reenterable seals, and said step of flowing labeled target molecules into said reaction chamber comprises the steps of:  
     piercing said seal of said inlet;  
     piercing said seal of said outlet; and  
     flowing said labeled target molecules from said inlet into said cavity and out through said outlet.

25

EP 0 695 941 A2

26

38. The method as recited in claim 35 further comprising the step of agitating said target molecules to facilitate reaction between said at least one probe array and said labeled target molecules.

5

39. The method as recited in claim 38 further comprising the step of identifying at least one location where said labeled target molecules are located on said at least one probe array.

10

40. The method as recited in claim 39 wherein said identifying step comprises the step of placing said package in a scanner, said scanner being capable of imaging said labeled target molecules on said at least one probe array.

15

41. The method as recited in claim 40 wherein said package comprises an alignment structure such that said package is placed in a desired position in said scanner.

20

42. The method as recited in claim 41 wherein said alignment structure comprises a non-flush edge.

43. The method as recited in claim 41 wherein said alignment structure comprises a plurality of alignment holes.

25

30

35

40

45

50

55

EP 0 695 941 A2

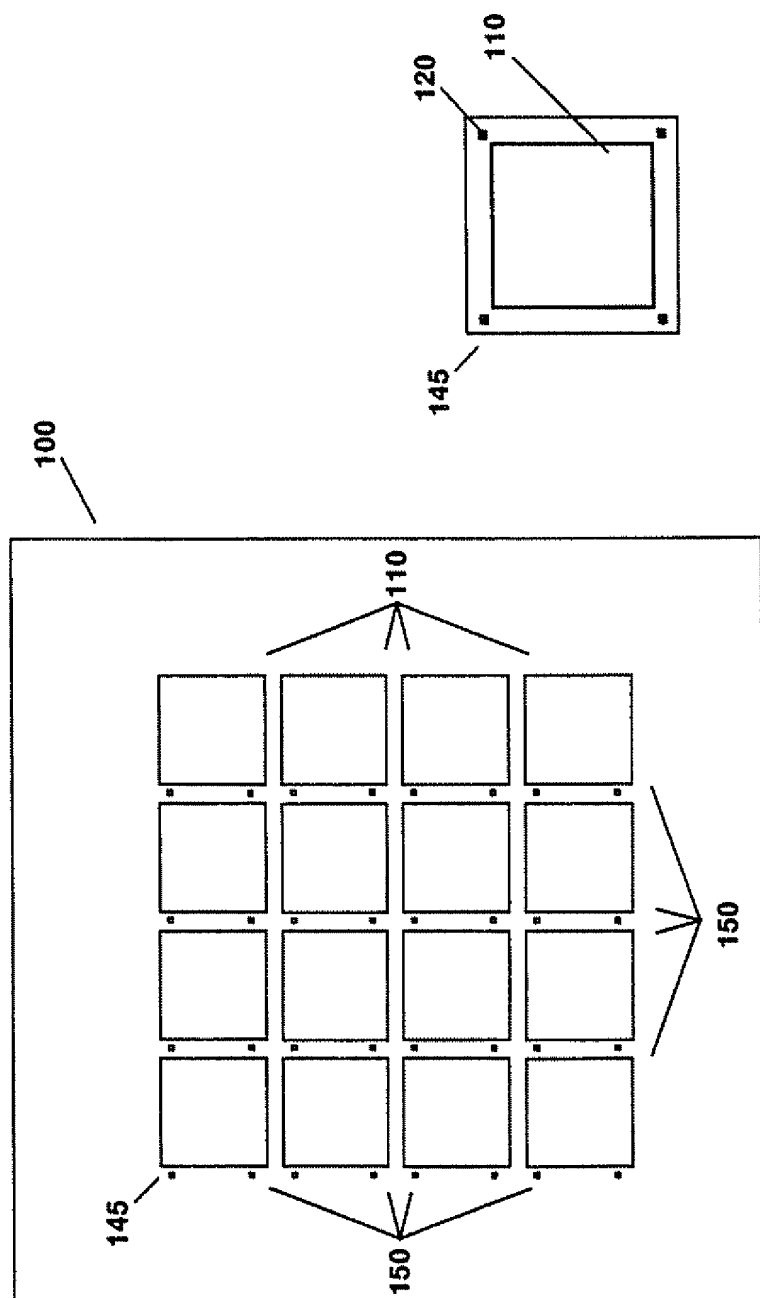


Figure 1b

Figure 1a

EP 0 695 941 A2

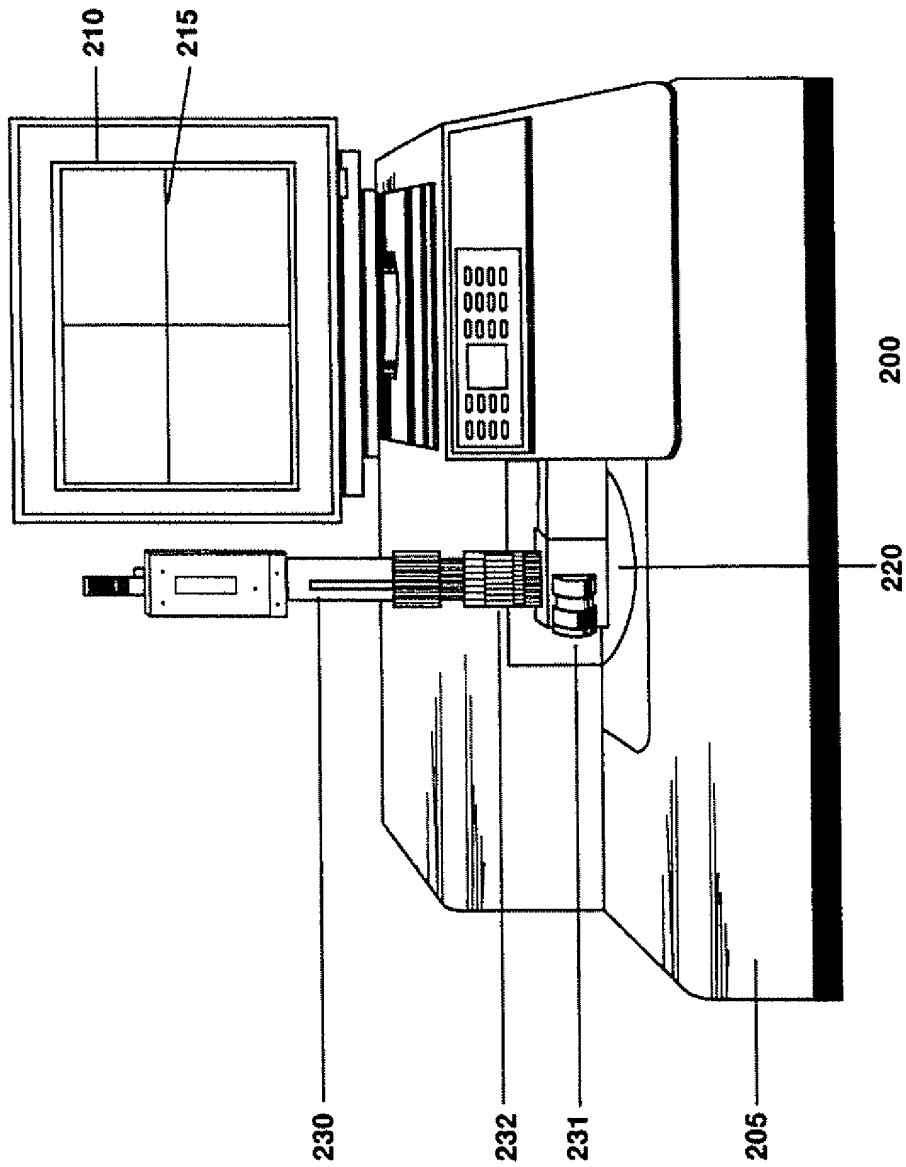


Figure 2a

EP 0 695 941 A2

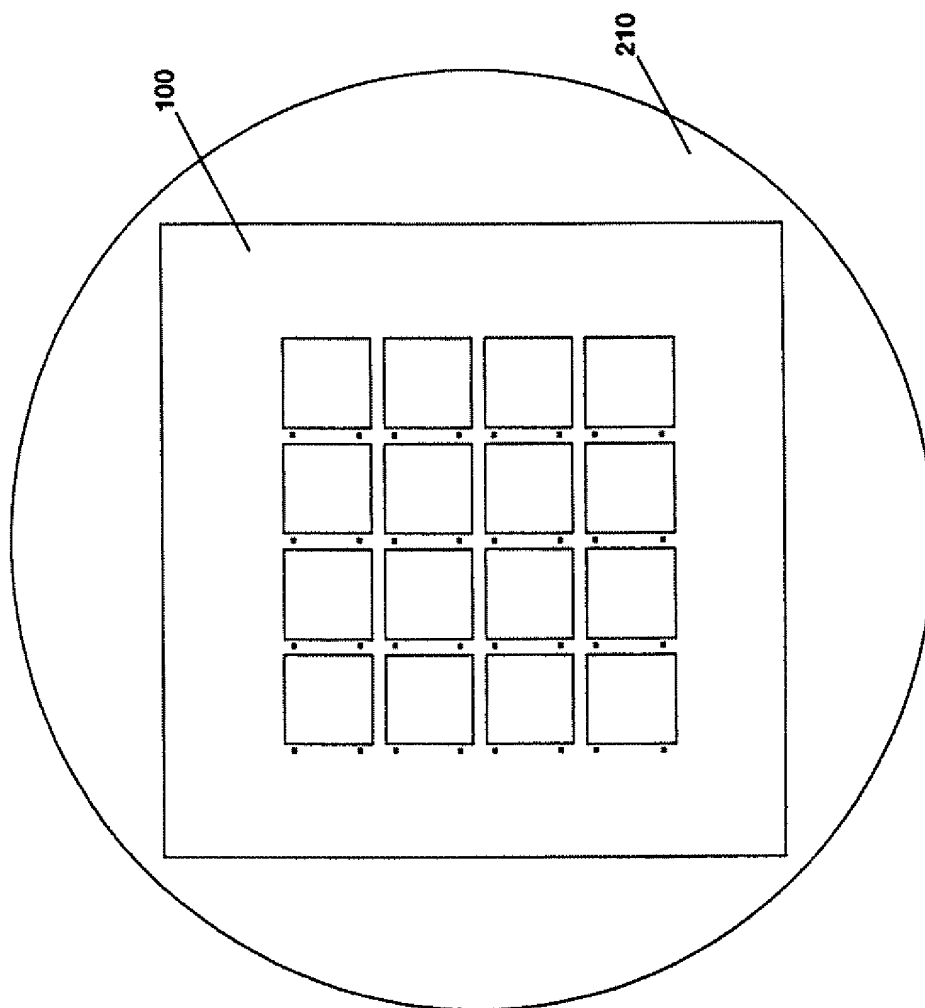


Figure 2b

EP 0 695 941 A2

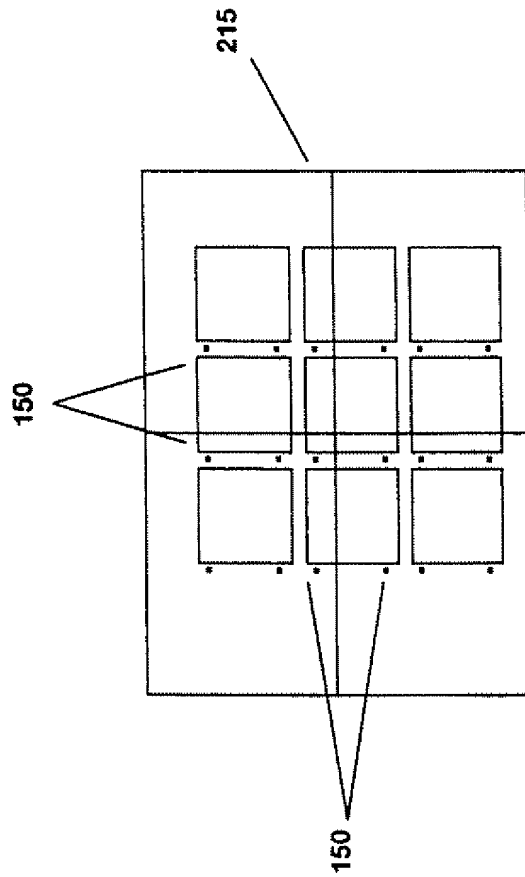


Figure 2c

EP 0 695 941 A2

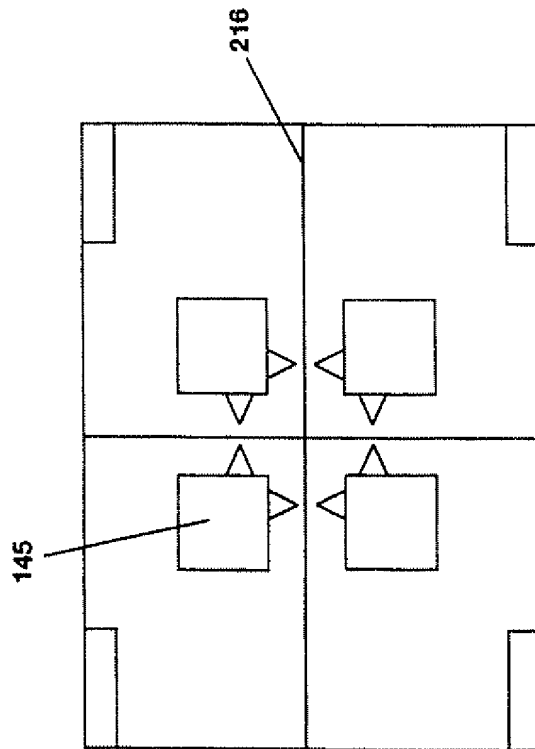


Figure 2d

EP 0 695 941 A2

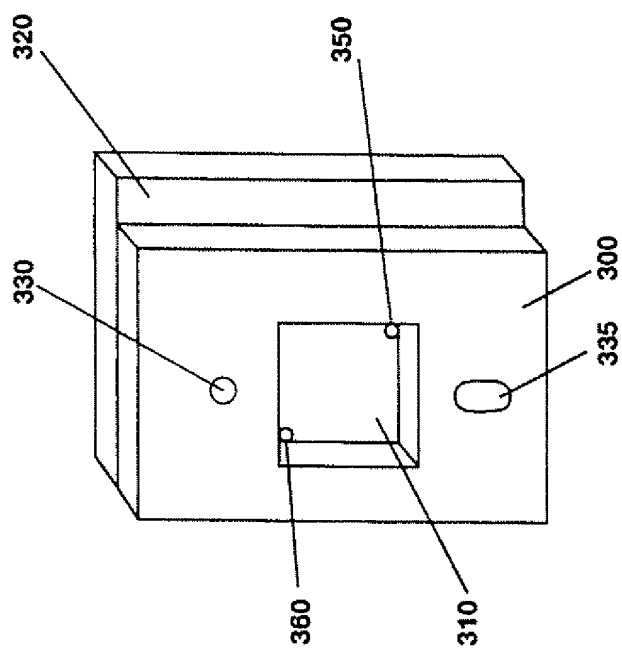


Figure 3

EP 0 695 941 A2

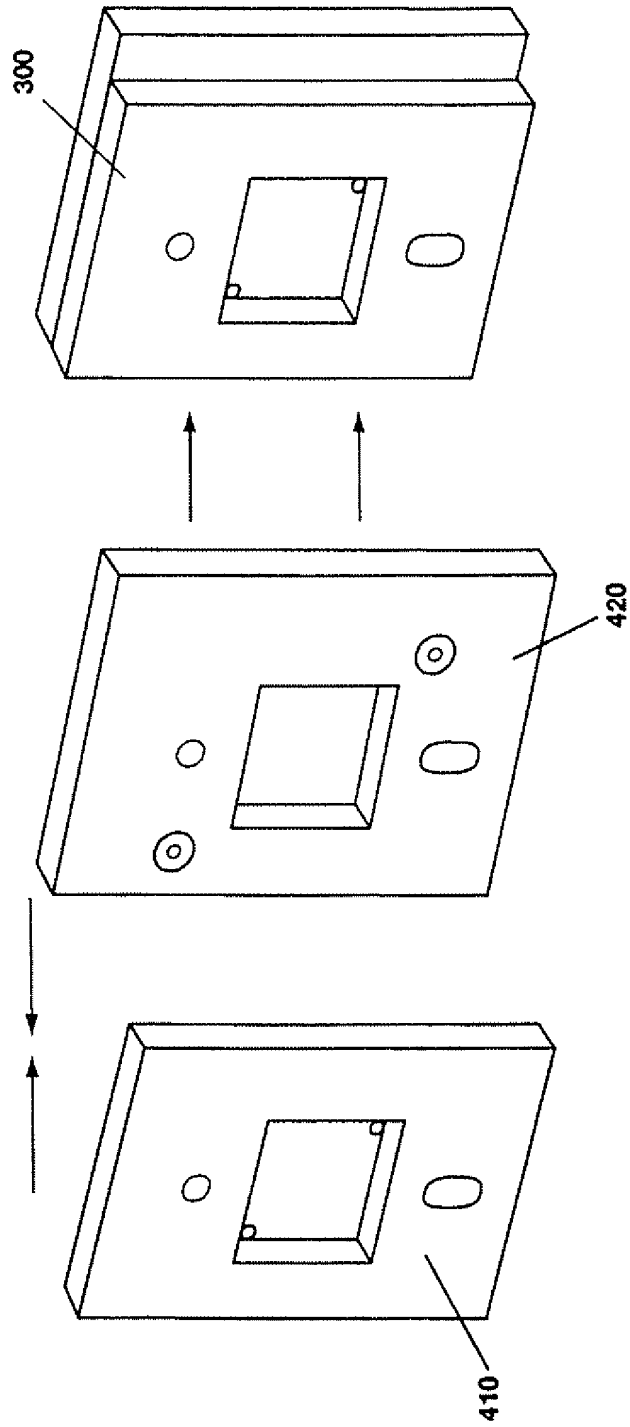


Figure 4

EP 0 695 941 A2

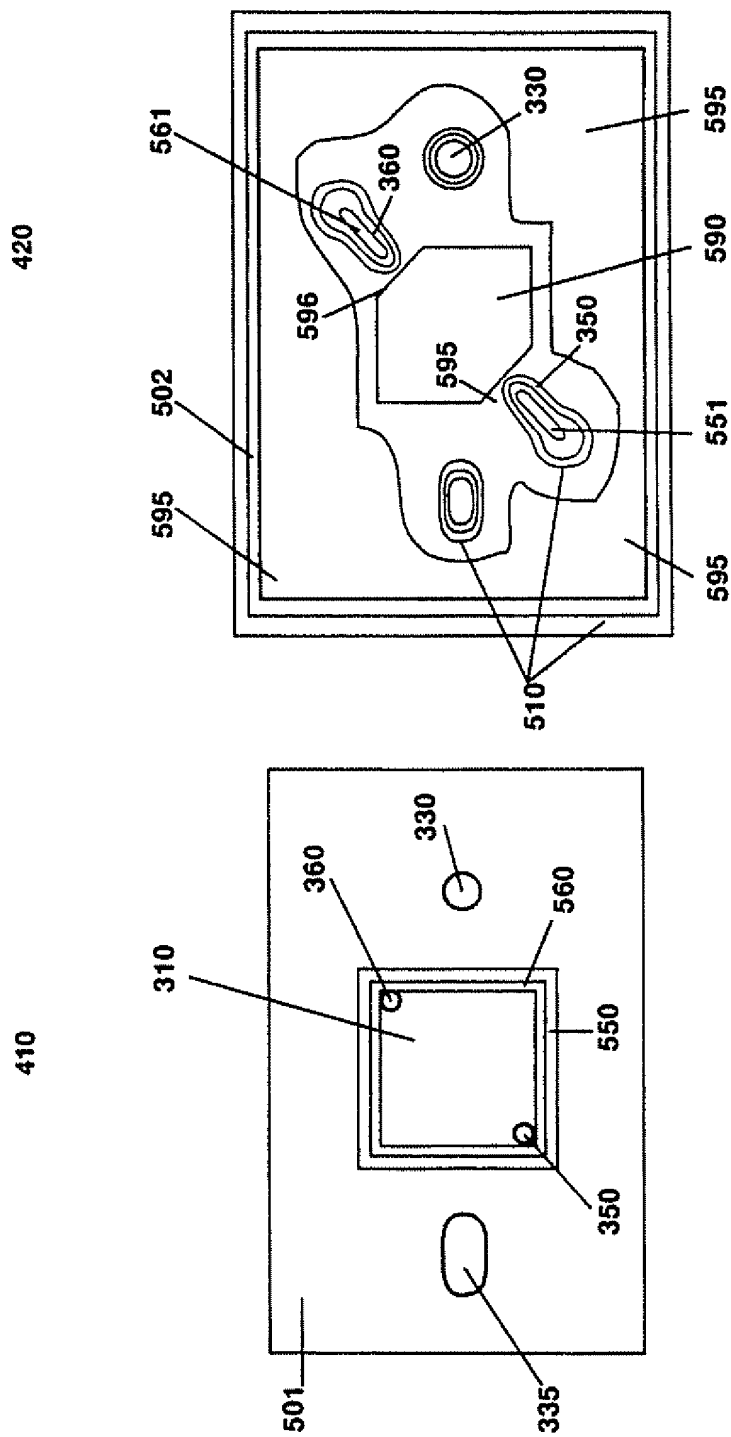


Figure 5b

Figure 5a

EP 0 695 941 A2

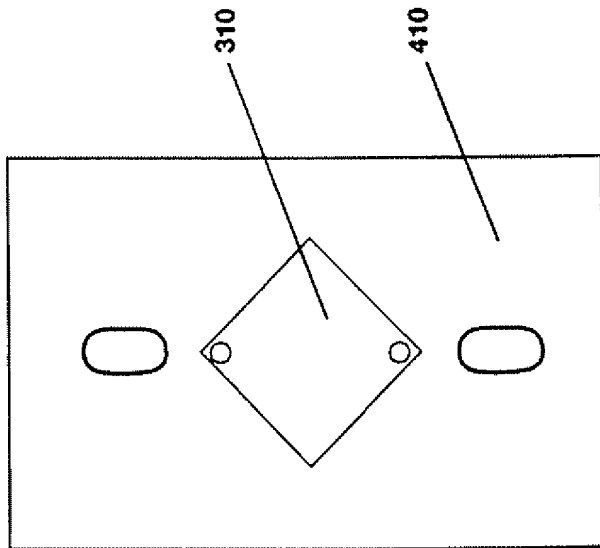


Figure 5c

EP 0 695 941 A2

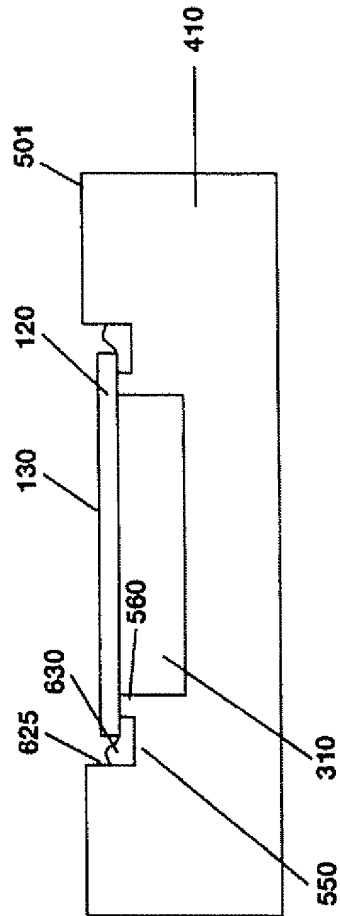


Figure 6

EP 0 695 941 A2

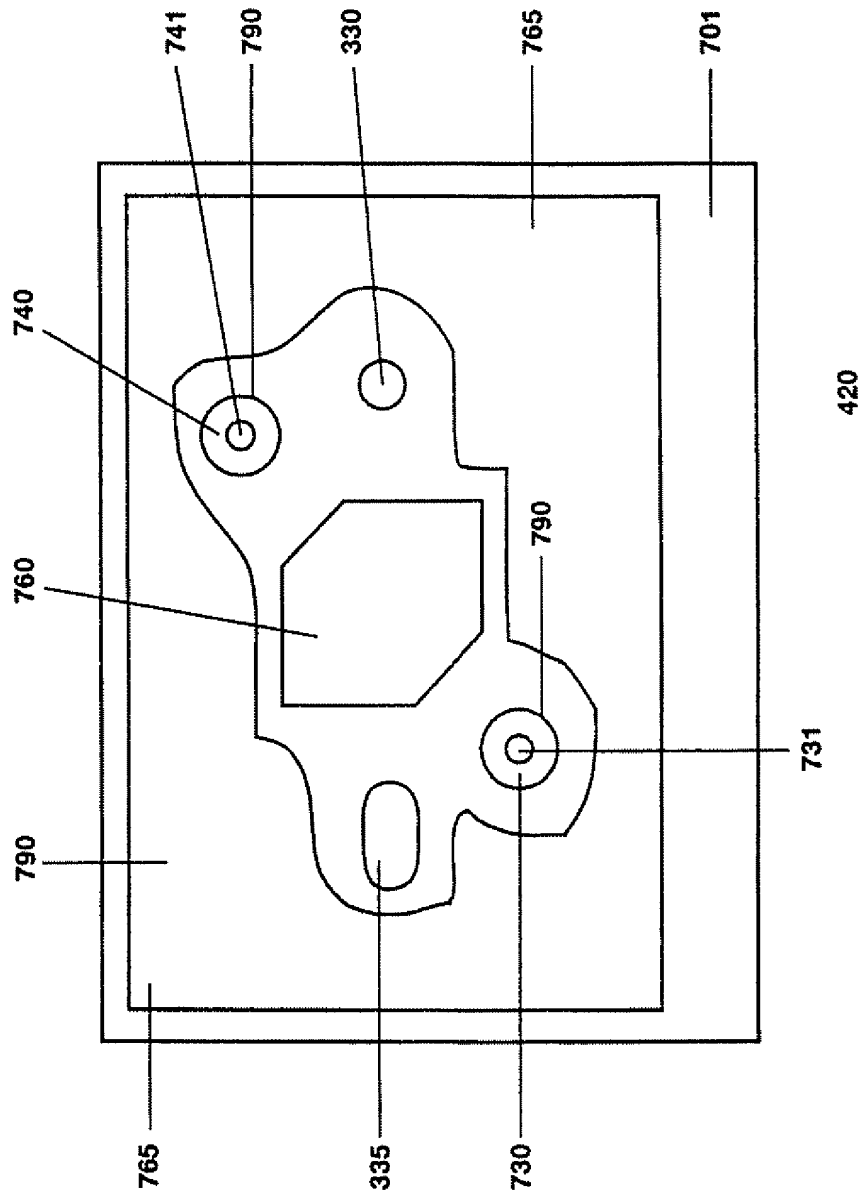


Figure 7

EP 0 695 941 A2

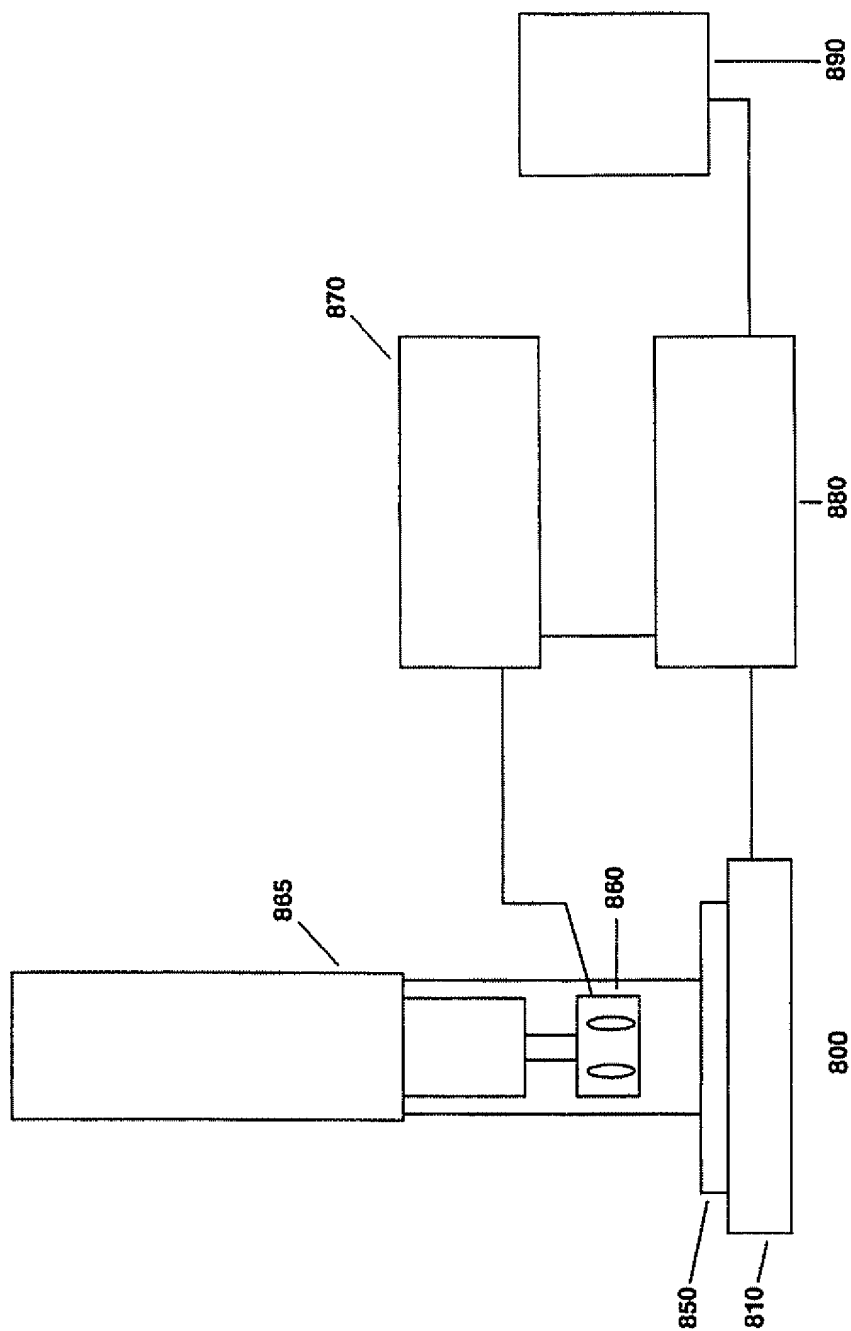


Figure 8a

EP 0 695 941 A2

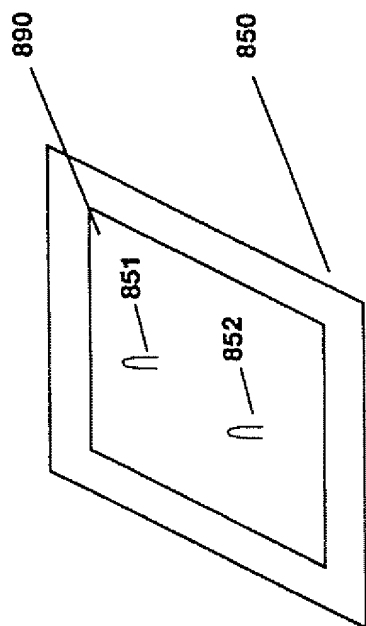


Figure 8b

EP 0 695 941 A2

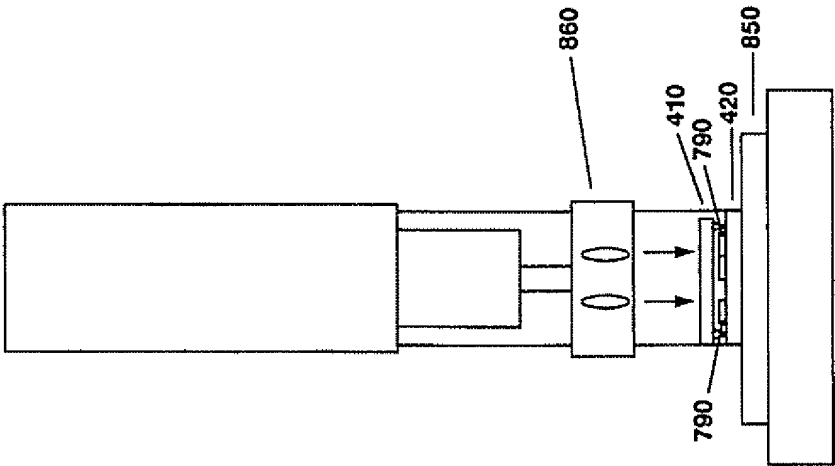


Figure 9a

EP 0 695 941 A2

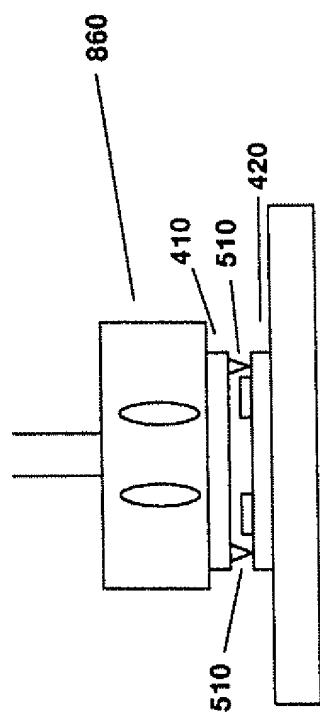


Figure 9b

EP 0 695 941 A2

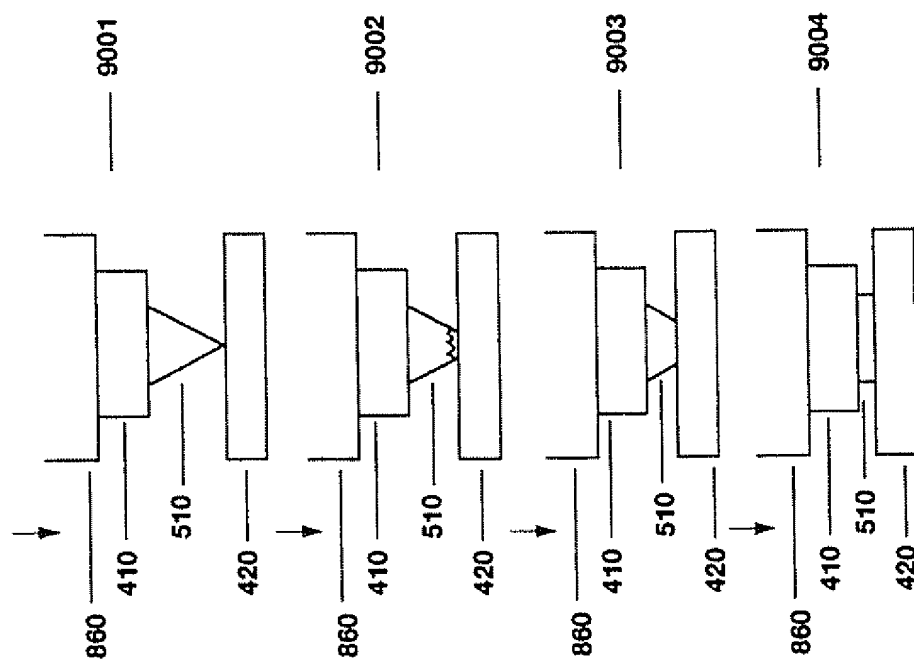


Figure 9c

EP 0 695 941 A2

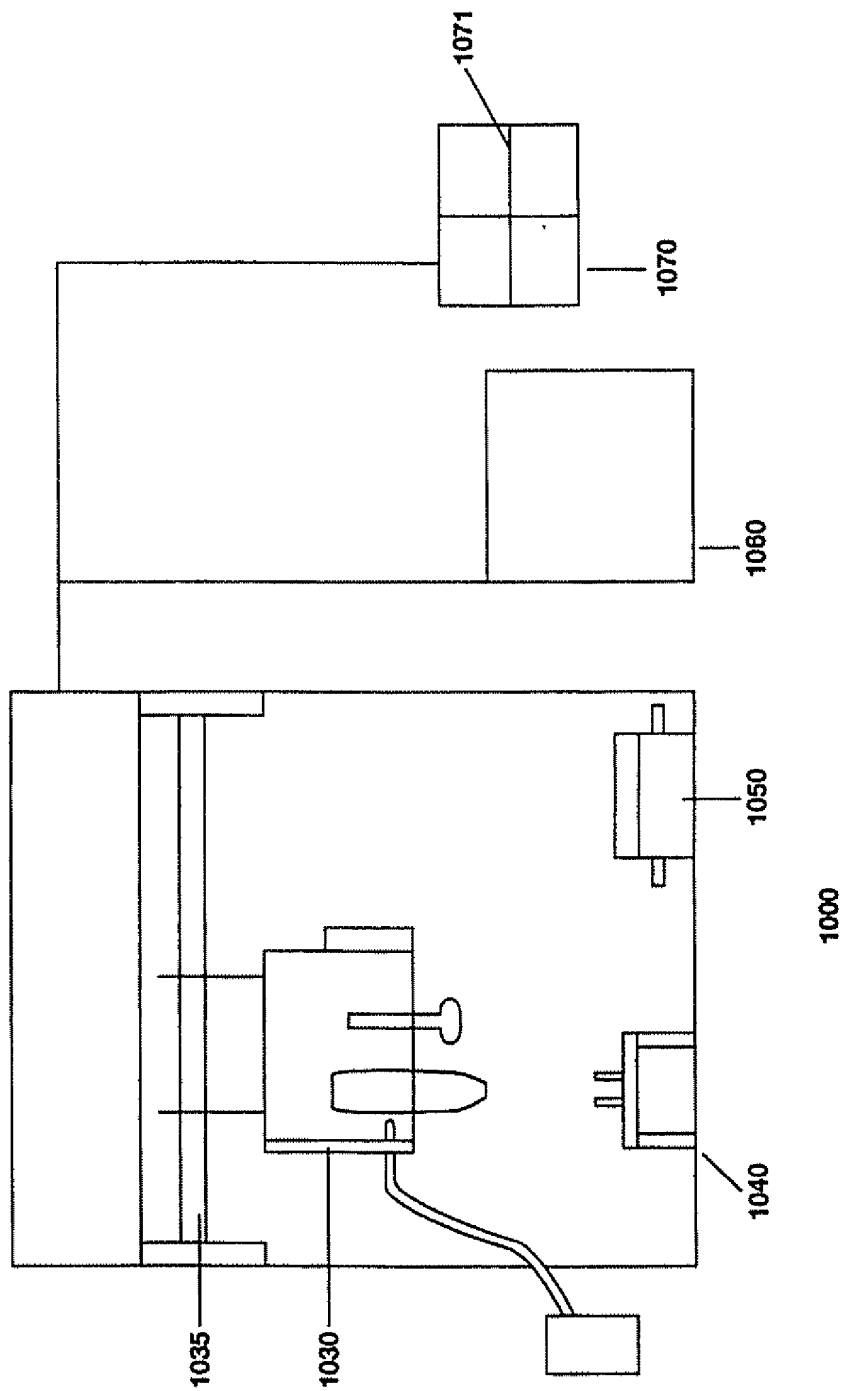


Figure 10

EP 0 695 941 A2

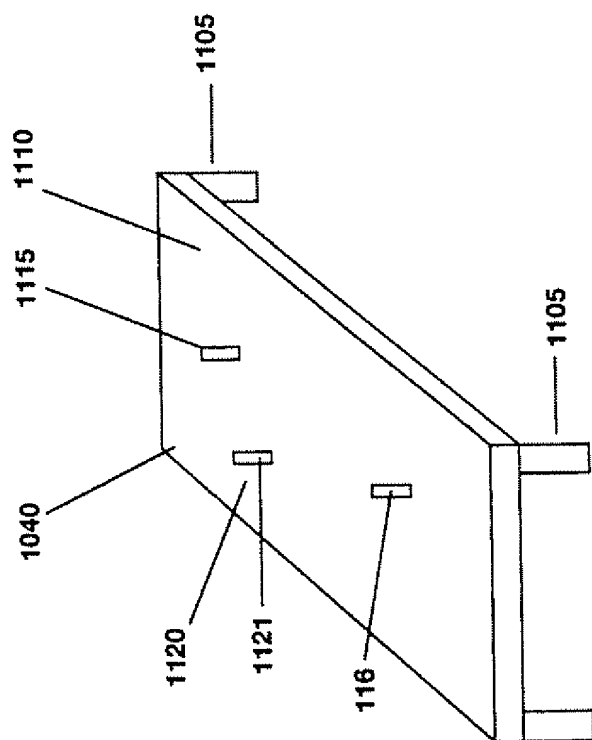


Figure 11

EP 0 695 941 A2

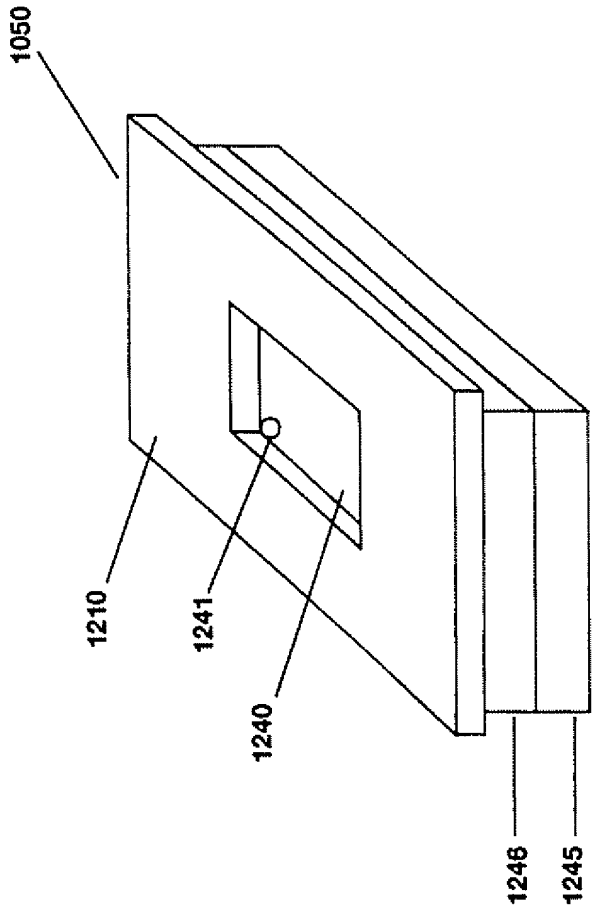


Figure 12a

EP 0 695 941 A2

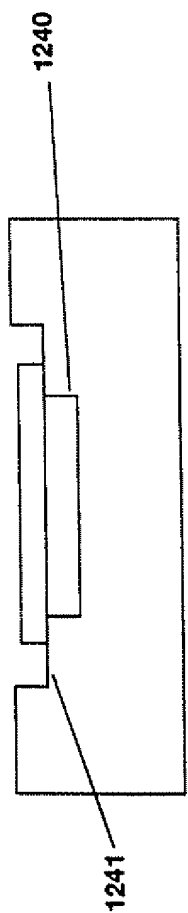


Figure 12b

EP 0 695 941 A2

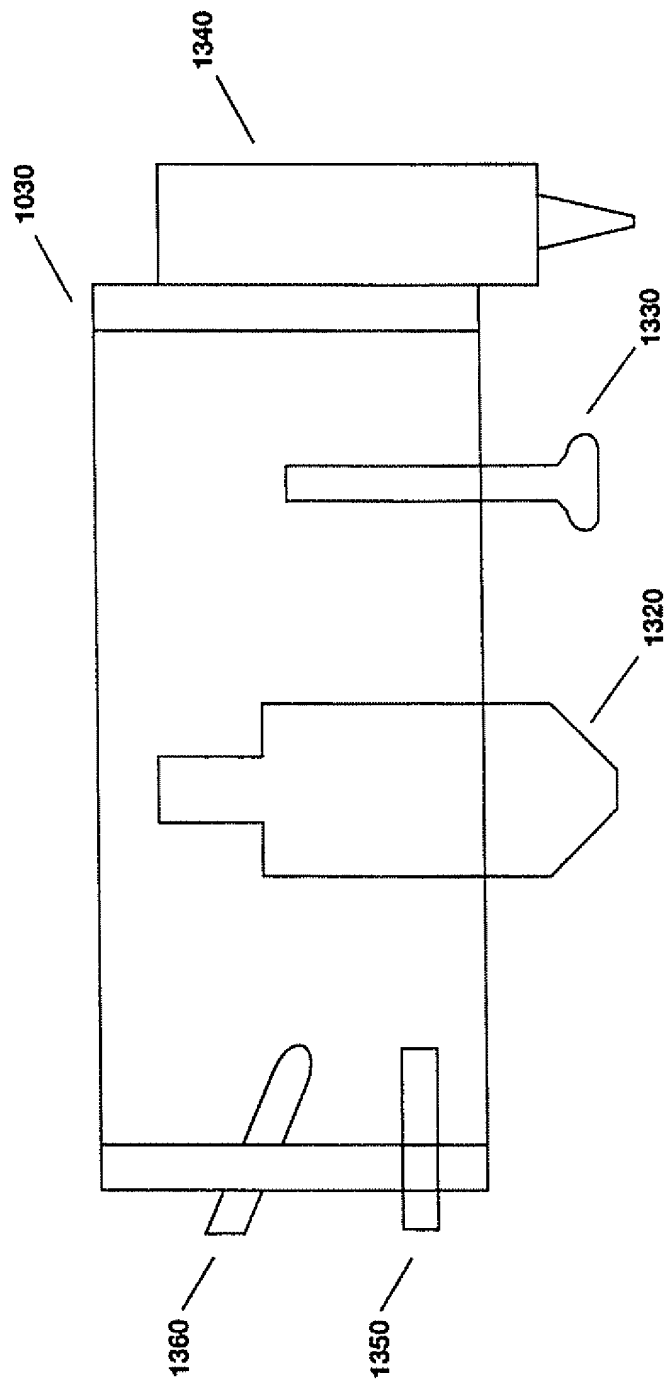


Figure 13

EP 0 695 941 A2

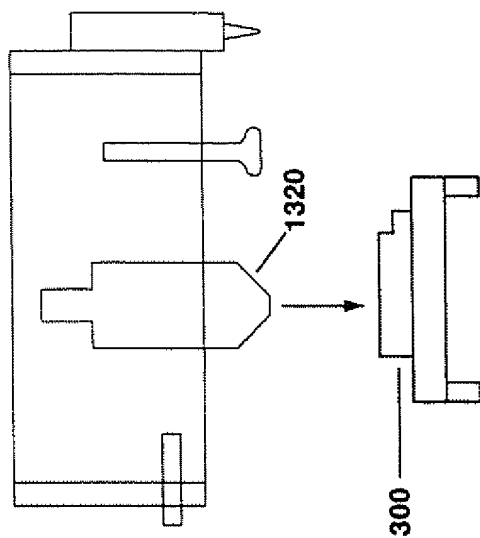


Figure 14a

EP 0 695 941 A2

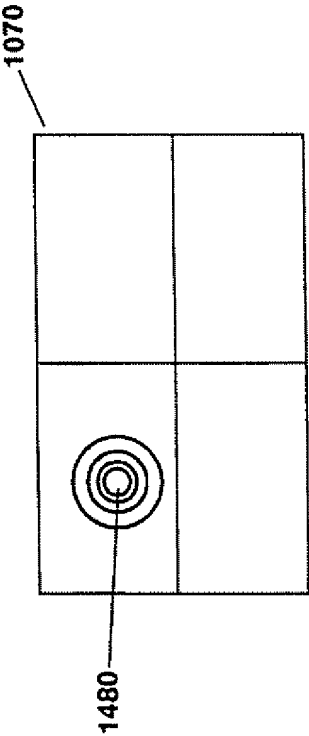


Figure 14b

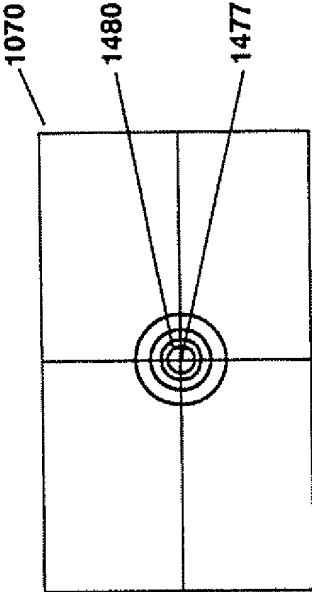


Figure 14c

## EP 0 695 941 A2

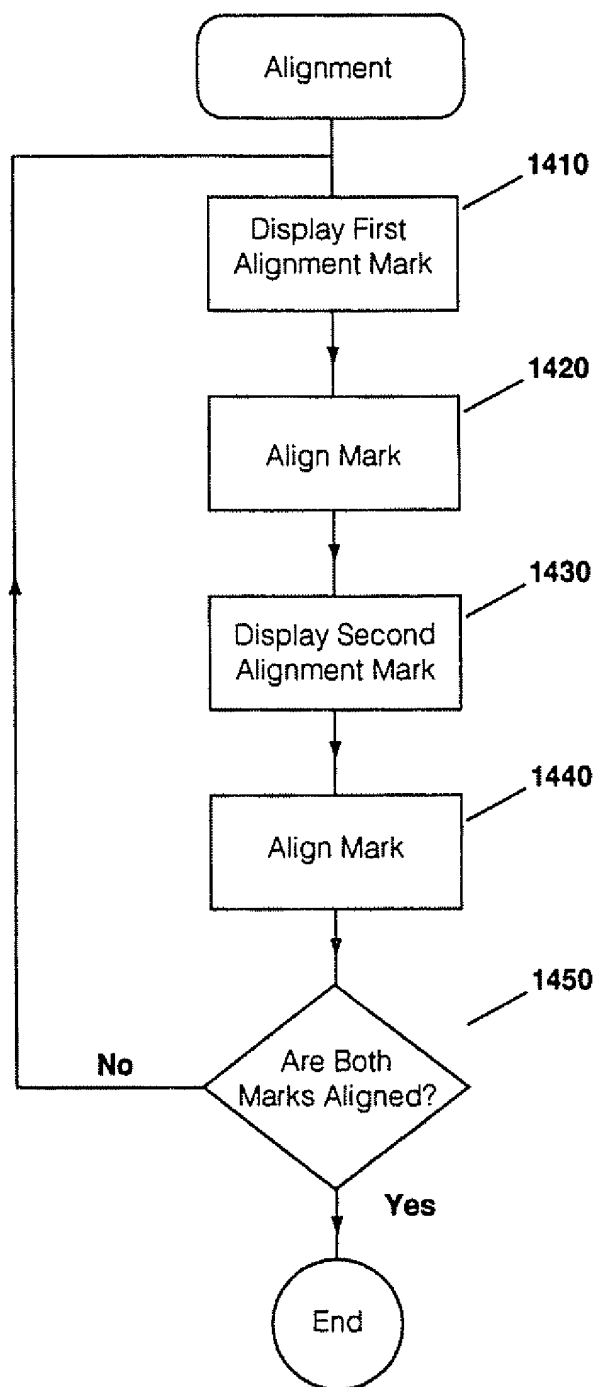


Figure 14d

EP 0 695 941 A2

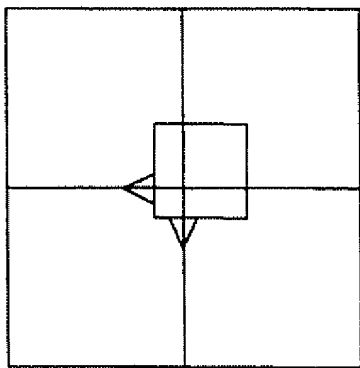


Figure 15e

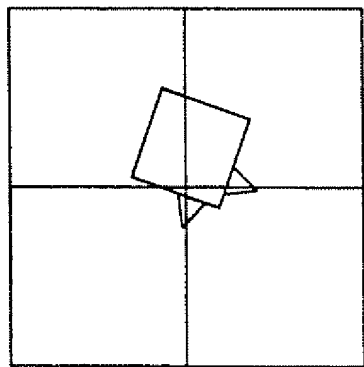


Figure 15b

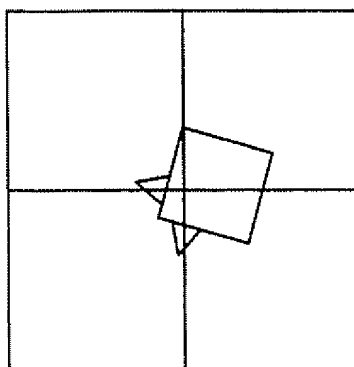


Figure 15d

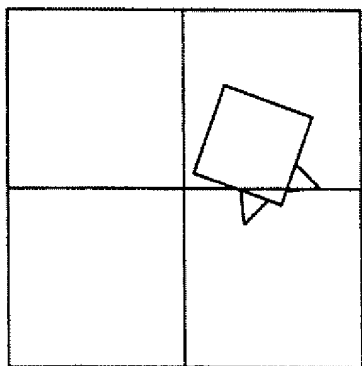


Figure 15a

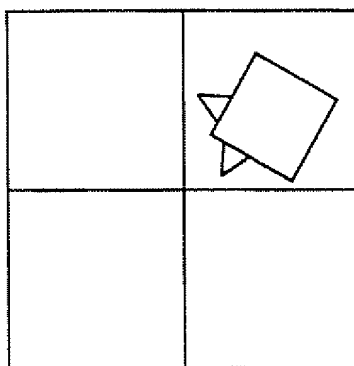


Figure 15c

## EP 0 695 941 A2

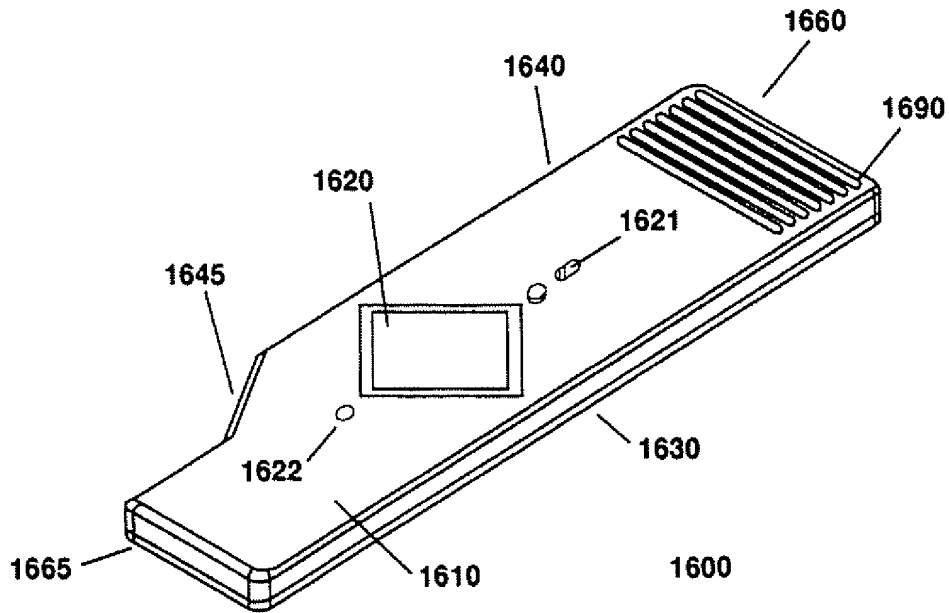


Figure 16a

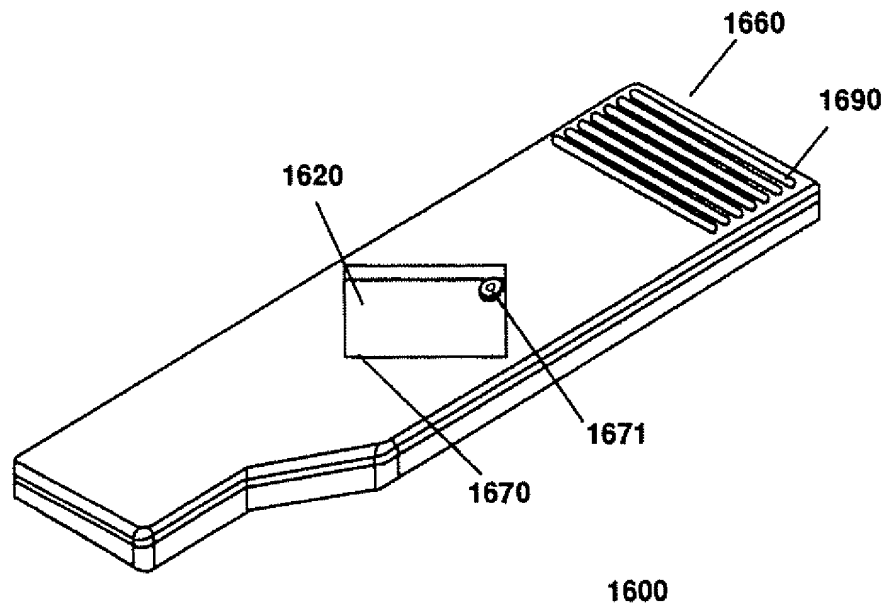


Figure 16b

EP 0 695 941 A2

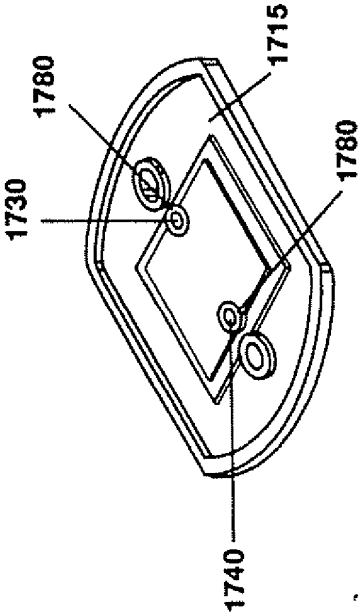
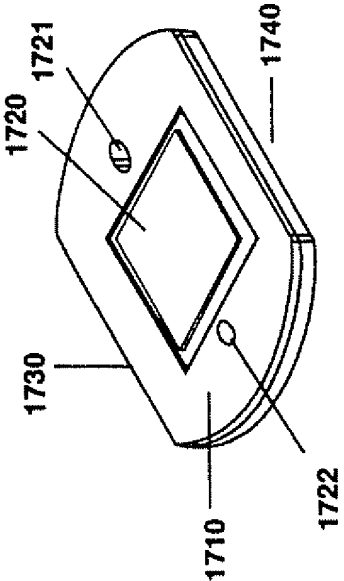


Figure 17b



1700

Figure 17a

EP 0 695 941 A2

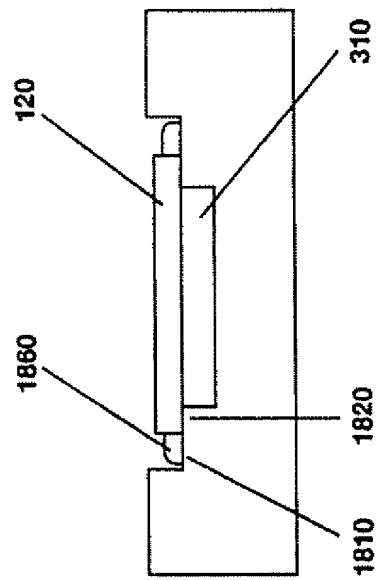


Figure 18

EP 0 695 941 A2

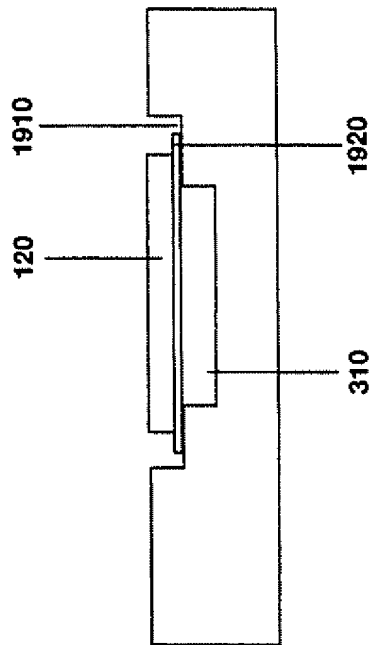


Figure19

EP 0 695 941 A2

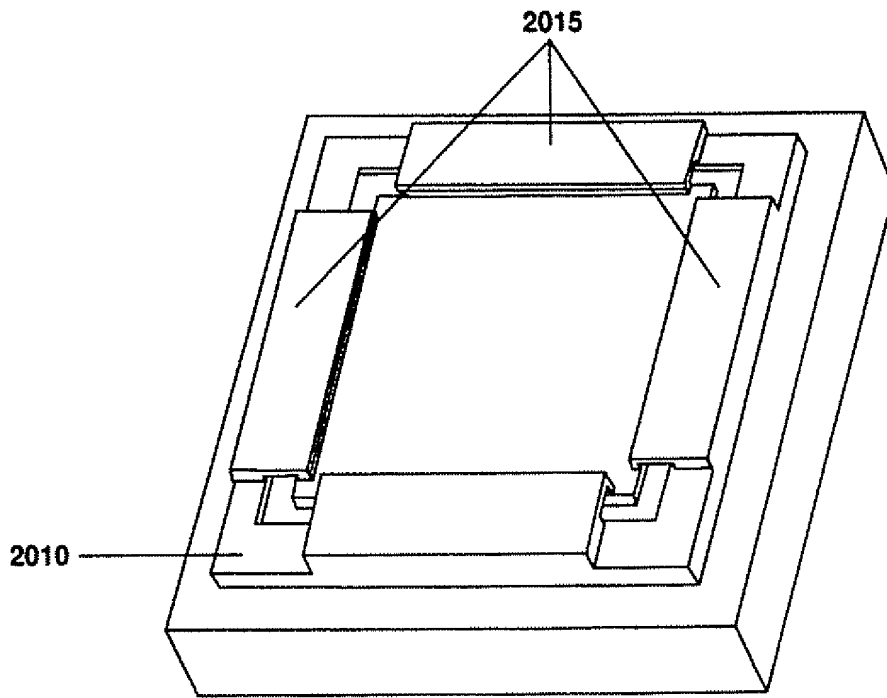


Figure 20a

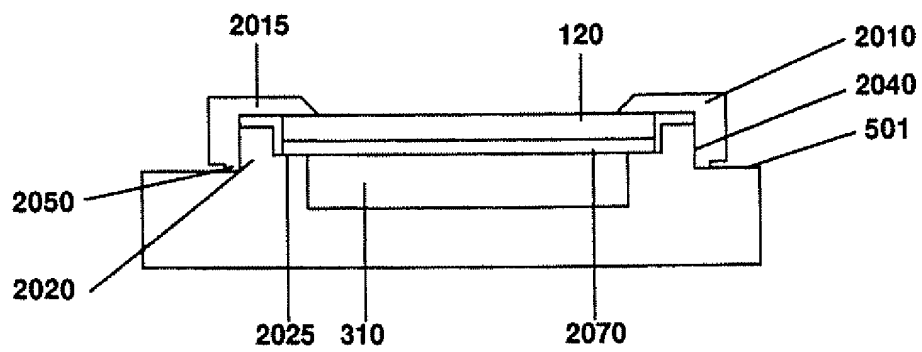


Figure 20b

EP 0 695 941 A2

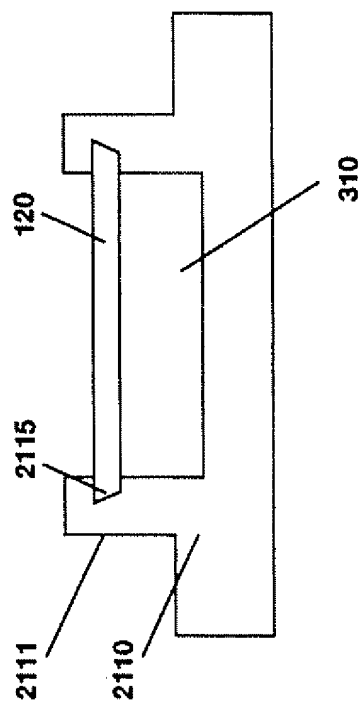


Figure 21

EP 0 695 941 A2

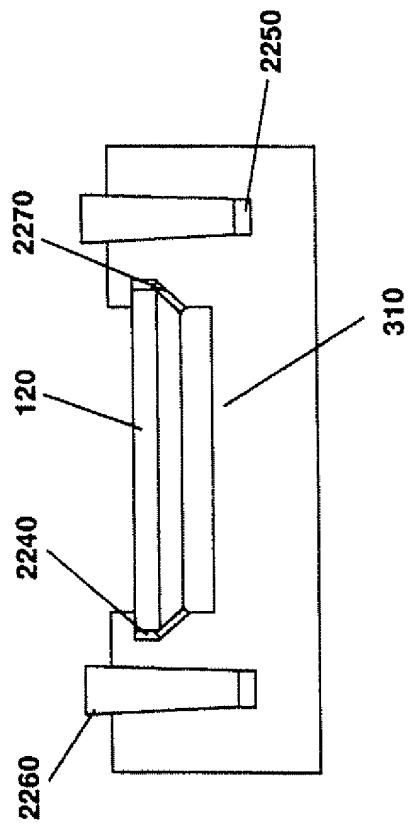


Figure 22

EP 0 695 941 A2

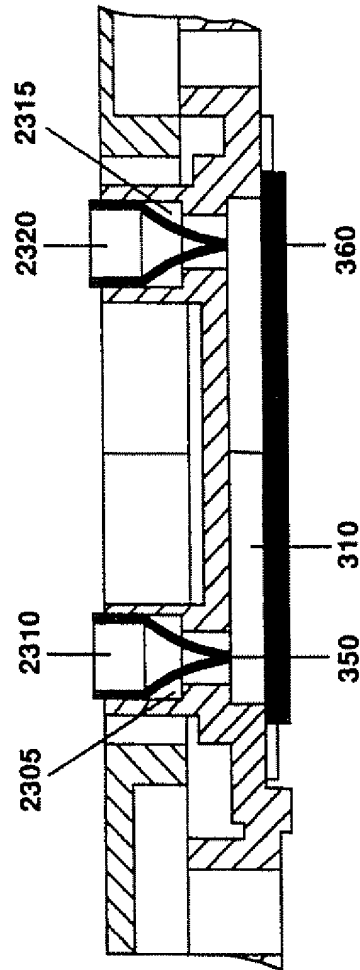


Figure 23

EP 0 695 941 A2

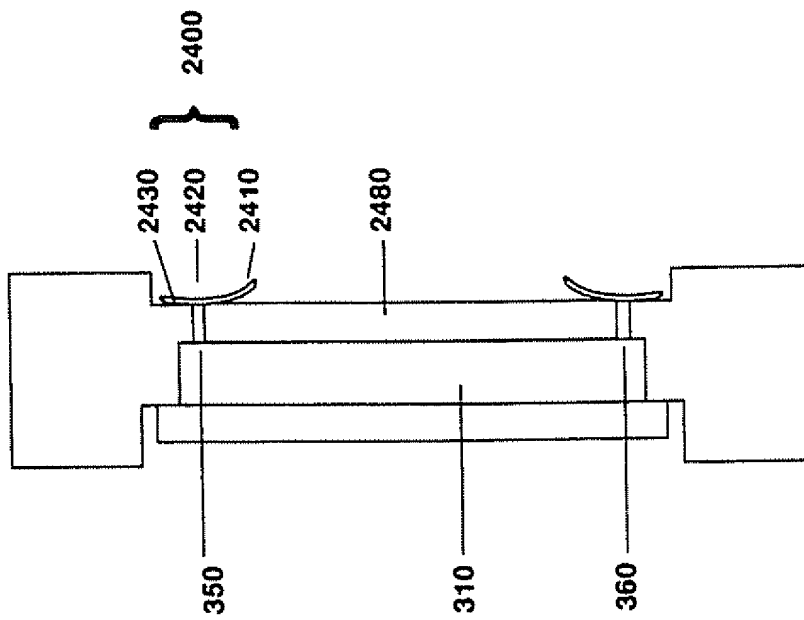


Figure 24

EP 0 695 941 A2

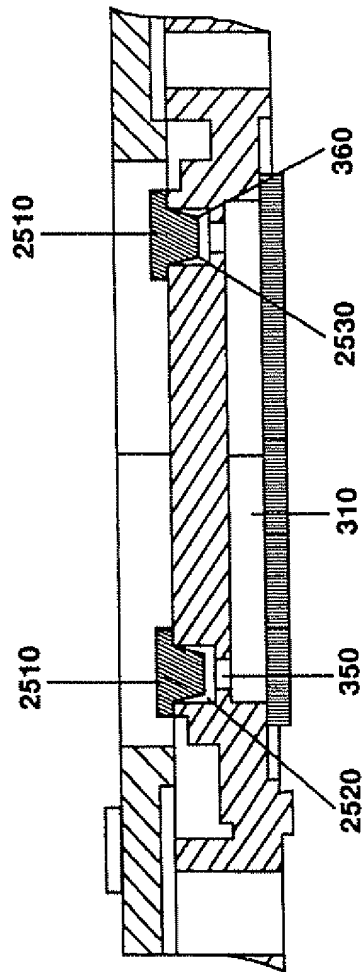


Figure 25

EP 0 695 941 A2

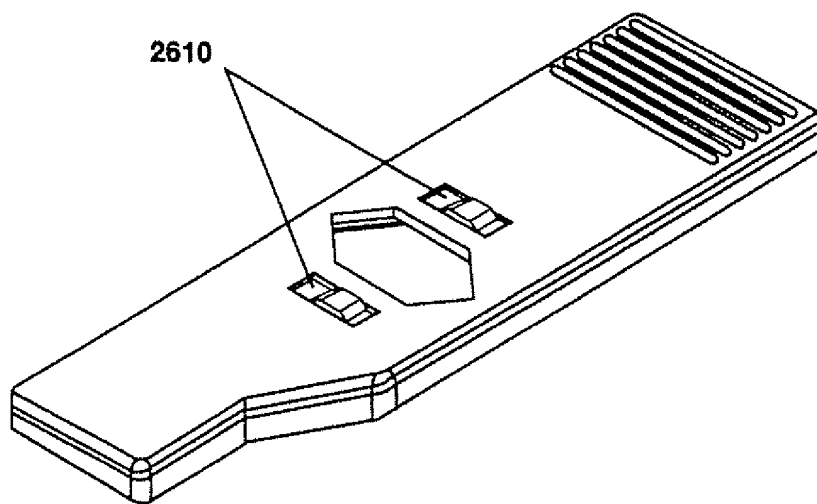


Figure 26a

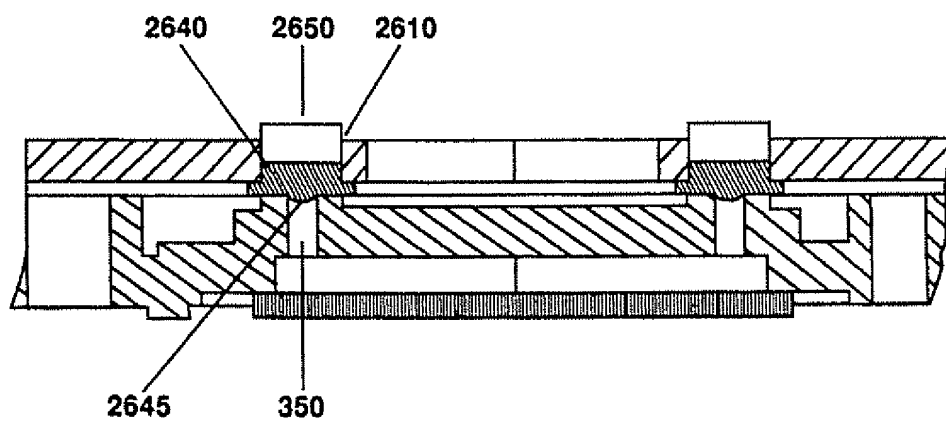


Figure 26b

EP 0 695 941 A2

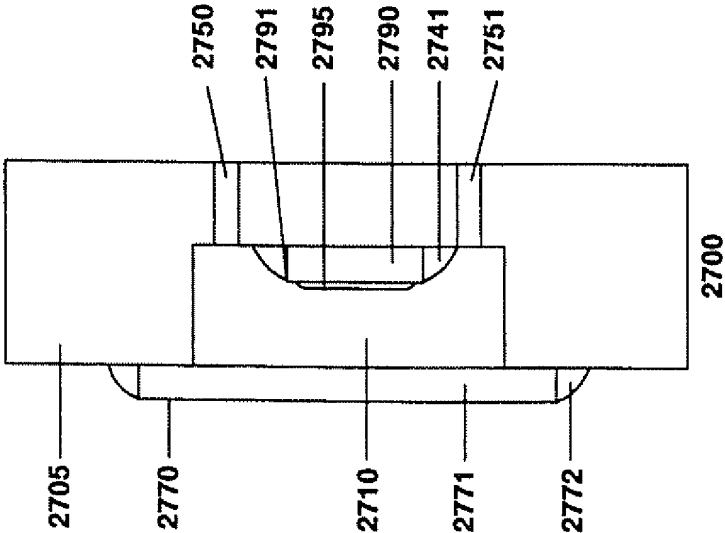


Figure 27b

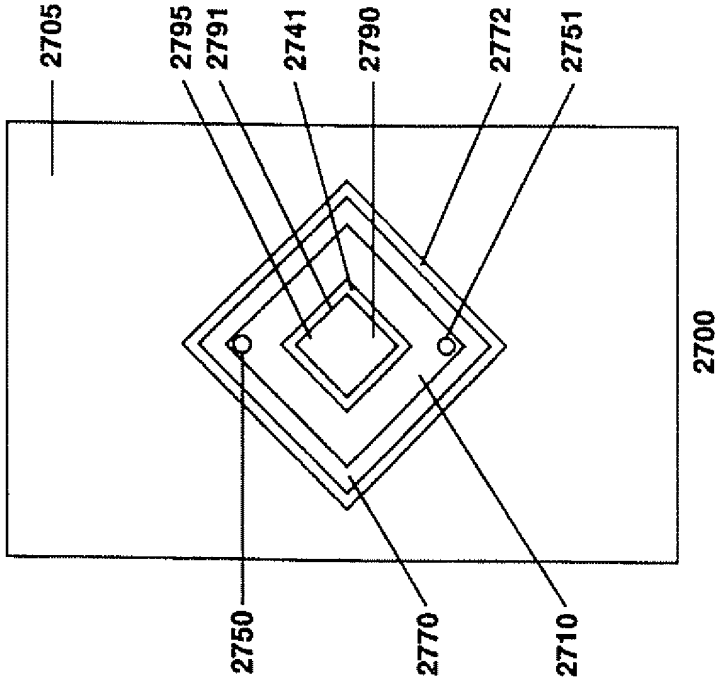


Figure 27a

EP 0 695 941 A2

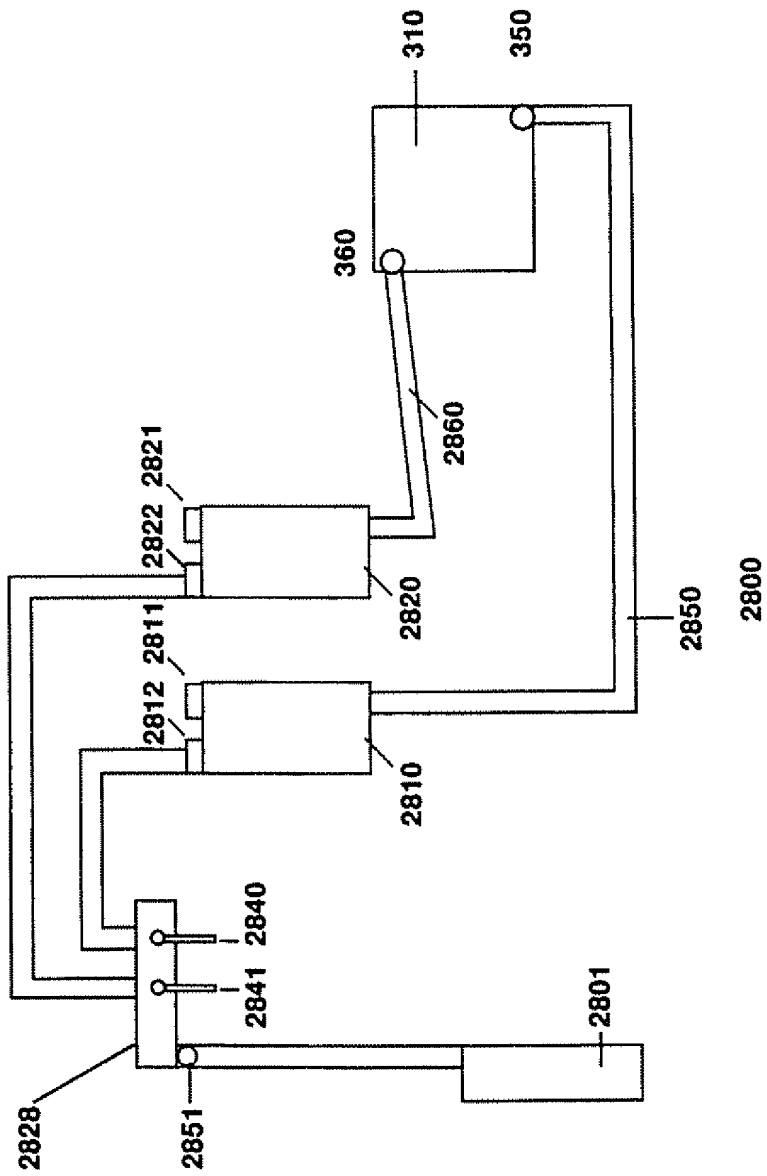


Figure 28

## EP 0 695 941 A2

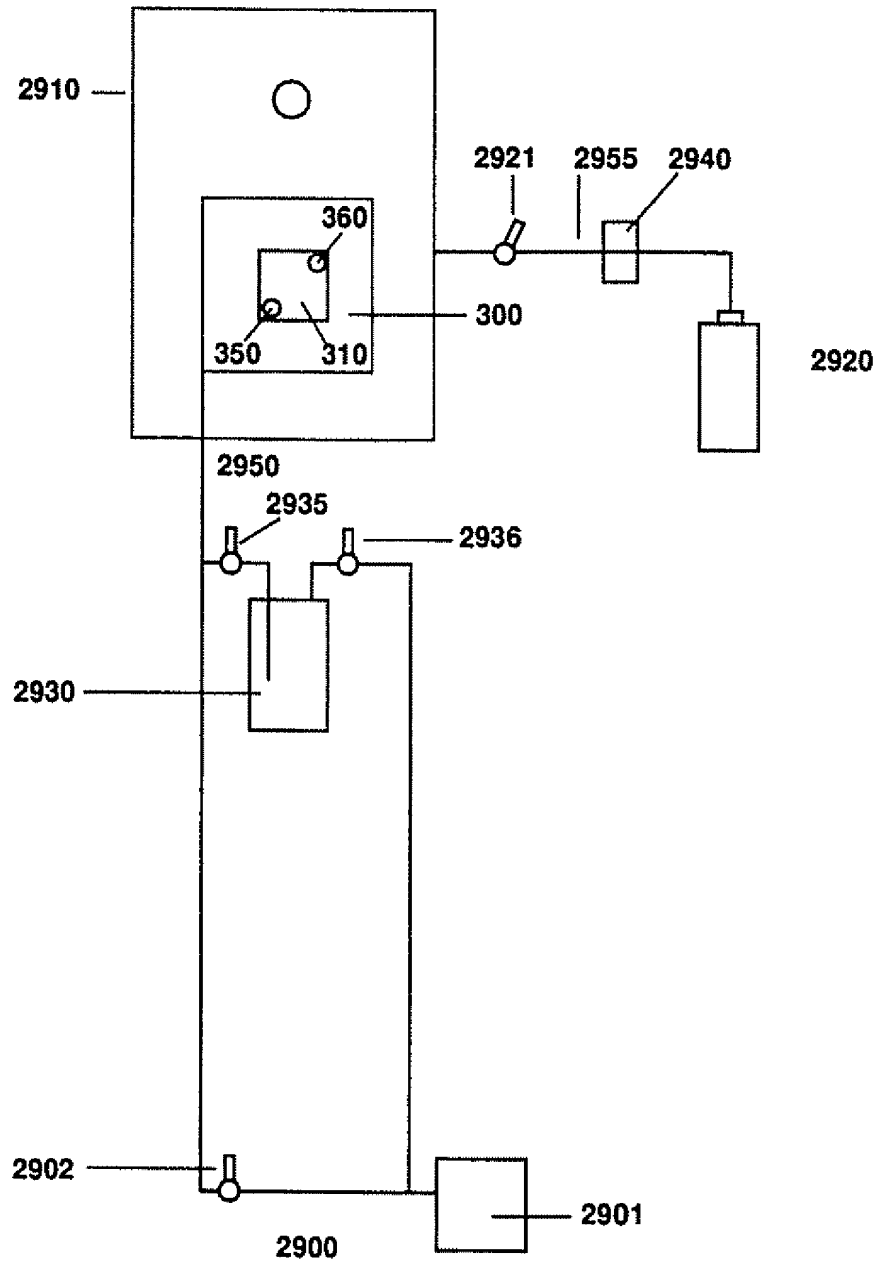


Figure 29

EP 0 695 941 A2

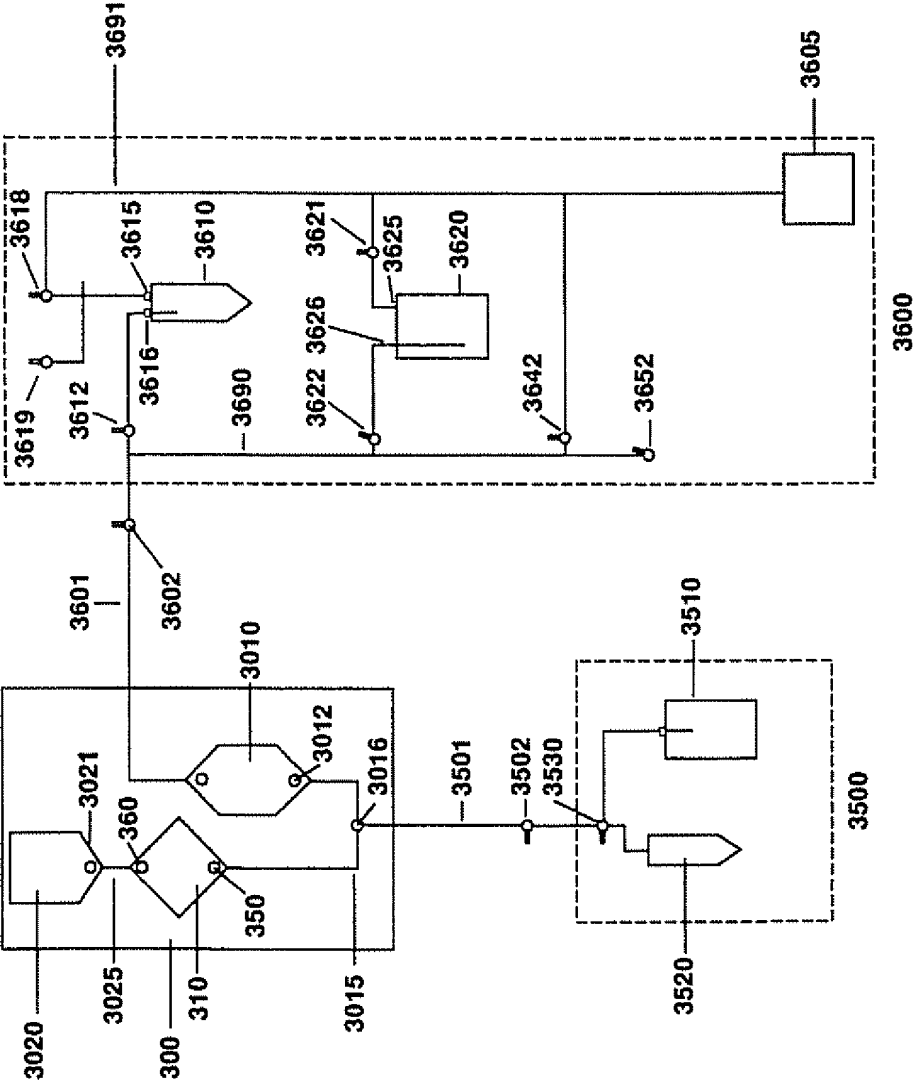


Figure 30

**PCT**WORLD INTELLECTUAL PROPERTY ORGANIZATION  
International Bureau

## INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

<b>(51) International Patent Classification <sup>5</sup> :</b> <b>C12Q 1/68, G01N 33/566, 33/48</b> <b>C07H 15/12</b>	<b>A1</b>	<b>(11) International Publication Number:</b> <b>WO 92/10588</b>  <b>(43) International Publication Date:</b> 25 June 1992 (25.06.92)
<b>(21) International Application Number:</b> PCT/US91/09226 <b>(22) International Filing Date:</b> 6 December 1991 (06.12.91)  <b>(30) Priority data:</b> 624,114 6 December 1990 (06.12.90) US  <b>(71) Applicant (for all designated States except US):</b> AFFYMAX TECHNOLOGIES N.V. [NL/NL]; De Ruyderkade 62, Curaçao (AN).  <b>(72) Inventors; and</b> <b>(75) Inventors/Applicants (for US only) :</b> FODOR, Stephen, P., A. [US/US]; 3863 Nathan Way, Palo Alto, CA 94303 (US). SOLAS, Dennis, W. [US/US]; 50 Gardenside Drive, #13, San Francisco, CA 94131 (US). DOWER, William, J. [US/US]; 761 Partridge Avenue, Menlo Park, CA 94025 (US).		<b>(74) Agents:</b> DUNN, Tracy, J. et al.; Townsend and Townsend, One Market Plaza, 2000 Steuart Tower, San Francisco, CA 94105 (US).  <b>(81) Designated States:</b> AT (European patent), AU, BE (European patent), CA, CH (European patent), DE (European patent), DK (European patent), ES (European patent), FR (European patent), GB (European patent), GR (European patent), IT (European patent), JP, LU (European patent), MC (European patent), NL (European patent), SE (European patent), US.  <b>Published</b> <i>With international search report.</i>
<b>(54) Title:</b> SEQUENCING BY HYBRIDIZATION OF A TARGET NUCLEIC ACID TO A MATRIX OF DEFINED OLIGONUCLEOTIDES  <b>(57) Abstract</b>  <p>The present invention provides methods and apparatus for sequencing, fingerprinting and mapping biological polymers, particularly polynucleotides. The methods make use of a plurality of positionally distinct sequence specific recognition reagents, such as polynucleotides. The apparatus employs a substrate comprising positionally distinct sequence specific recognition reagents, such as polynucleotides, which are preferably localized at high densities. The methods and apparatus of the present invention can be used for determining the sequence of polynucleotides, mapping polynucleotides, and developing polynucleotide fingerprints. Polynucleotide fingerprints can be used for identifying individuals, tissue samples, pathological conditions, genetic diseases, infectious diseases, and other applications. Polynucleotide fingerprints can also be used for classification of biological samples, including taxonomy, and to characterize their sources. The invention also provides polynucleotide mapping, fingerprinting, and sequencing as valuable laboratory research tools for use in biological investigations.</p>		

**FOR THE PURPOSES OF INFORMATION ONLY**

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	ES	Spain	MG	Madagascar
AU	Australia	FI	Finland	ML	Mali
BB	Barbados	FR	France	MN	Mongolia
BE	Belgium	GA	Gabon	MR	Mauritania
BF	Burkina Faso	GB	United Kingdom	MW	Malawi
BG	Bulgaria	GN	Guinea	NL	Netherlands
BJ	Benin	GR	Greece	NO	Norway
BR	Brazil	HU	Hungary	PL	Poland
CA	Canada	IT	Italy	RO	Romania
CF	Central African Republic	JP	Japan	SD	Sudan
CG	Congo	KP	Democratic People's Republic of Korea	SE	Sweden
CH	Switzerland	KR	Republic of Korea	SN	Senegal
CI	Côte d'Ivoire	LI	Liechtenstein	SU <sup>+</sup>	Soviet Union
CM	Cameroon	LK	Sri Lanka	TD	Chad
CS	Czechoslovakia	LU	Luxembourg	TC	Togo
DE*	Germany	MC	Monaco	US	United States of America
DK	Denmark				

+ Any designation of "SU" has effect in the Russian Federation. It is not yet known whether any such designation has effect in other States of the former Soviet Union.

WO 92/10588

PCT/US91/09226

1

5                   SEQUENCING BY HYBRIDIZATION OF A TARGET NUCLEIC ACID  
                  TO A MATRIX OF DEFINED OLIGONUCLEOTIDES

                  BACKGROUND OF THE INVENTION

                  The present invention relates to the sequencing,  
fingerprinting, and mapping of polymers, particularly  
10 biological polymers. The inventions may be applied, for  
example, in the sequencing, fingerprinting, or mapping of  
polynucleotides.

                  The relationship between structure and function of  
macromolecules is of fundamental importance in the  
15 understanding of biological systems. These relationships are  
important to understanding, for example, the functions of  
enzymes, structural proteins, and signalling proteins, ways in  
which cells communicate with each other, as well as mechanisms  
of cellular control and metabolic feedback.

20                   Genetic information is critical in continuation of  
life processes. Life is substantially informationally based  
and its genetic content controls the growth and reproduction of  
the organism and its complements. Polypeptides, which are  
critical features of all living systems, are encoded by the  
25 genetic material of the cell. In particular, the properties of  
enzymes, functional proteins, and structural proteins are  
determined by the sequence of amino acids which make them up.  
As structure and function are integrally related, many  
biological functions may be explained by elucidating the  
30 underlying the structural features which provide those  
functions. For this reason, it has become very important to  
determine the genetic sequences of nucleotides which encode the  
enzymes, structural proteins, and other effectors of biological  
functions. In addition to segments of nucleotides which encode  
35 polypeptides, there are many nucleotide sequences which are  
involved in control and regulation of gene expression.

                  The human genome project is directed toward  
determining the complete sequence the genome of the human  
organism. Although such a sequence would not correspond to the

WO 92/10588

PCT/US91/09226

2

sequence of any specific individual, it would provide significant information as to the general organization and specific sequences contained within segments from particular individuals. It would also provide mapping information which is very useful for further detailed studies. However, the need for highly rapid, accurate, and inexpensive sequencing technology is nowhere more apparent than in a demanding sequencing project such as this. To complete the sequencing of a human genome would require the determination of approximately  $3 \times 10^9$ , or 3 billion base pairs.

The procedures typically used today for sequencing include the Sanger dideoxy method, see, e.g., Sanger et al. (1977) Proc. Natl. Acad. Sci. USA, 74:5463-5467, or the Maxam and Gilbert method, see, e.g., Maxam et al., (1980) Methods in Enzymology, 65:499-559. The Sanger method utilizes enzymatic elongation procedures with chain terminating nucleotides. The Maxam and Gilbert method uses chemical reactions exhibiting specificity of reaction to generate nucleotide specific cleavages. Both methods require a practitioner to perform a large number of complex manual manipulations. These manipulations usually require isolating homogeneous DNA fragments, elaborate and tedious preparing of samples, preparing a separating gel, applying samples to the gel, electrophoresing the samples into this gel, working up the finished gel, and analyzing the results of the procedure.

Thus, a less expensive, highly reliable, and labor efficient means for sequencing biological macromolecules is needed. A substantial reduction in cost and increase in speed of nucleotide sequencing would be very much welcomed. In particular, an automated system would improve the reproducibility and accuracy of procedures. The present invention satisfies these and other needs.

#### SUMMARY OF THE INVENTION

The present invention provides improved methods useful for de novo sequencing of an unknown polymer sequence, for verification of known sequences, for fingerprinting polymers, and for mapping homologous segments within a

WO 92/10588

PCT/US91/09226

3

sequence. By reducing the number of manual manipulations required and automating most of the steps, the speed, accuracy, and reliability of these procedures are greatly enhanced.

The production of a substrate having a matrix of positionally defined regions with attached reagents exhibiting known recognition specificity can be used for the sequence analysis of a polymer. Although most directly applicable to sequencing, the present invention is also applicable to fingerprinting, mapping, and general screening of specific interactions.

The present invention also provides a means to automate sequencing manipulations. The automation of the substrate production method and of the scan and analysis steps minimizes the need for human intervention. This simplifies the tasks and promotes reproducibility.

The present invention provides a composition comprising a plurality of positionally distinguishable sequence specific reagents attached to a solid substrate, which reagents are capable of specifically binding to a predetermined subunit sequence of a preselected multi-subunit length having at least three subunits, said reagents representing substantially all possible sequences of said preselected length. In some embodiments, the subunit sequence is a polynucleotide sequence. In other embodiments, the specific reagent is an oligonucleotide of at least about five nucleotides, preferably at least eight nucleotides, more preferably at least 12 nucleotides. Usually the specific reagents are all attached to a single solid substrate, and the reagents comprise at least 3000 different sequences. In other embodiments, the reagents represents at least about 25% of the possible subsequences of said preselected length. Usually, the reagents are localized in regions of the substrate having a density of at least 25 regions per square centimeter, and often the substrate has a surface area of less than about 4 square centimeters.

The present invention also provides methods for analyzing a sequence of a polynucleotide, said method comprising the step of:

WO 92/10588

PCT/US91/09226

4

- a) exposing said polynucleotide to a composition as described.

It also provides useful methods for identifying or comparing a target sequence with a reference (i.e., fingerprinting), said method comprising the step of:

- 5 a) exposing said target sequence to a composition as described;
- b) determining the pattern of positions of the reagents which specifically interact with the target sequence; and
- 10 c) comparing the pattern with the pattern exhibited by the reference when exposed to the composition.

By way of example and not limitation, such fingerprinting methods may be used for personal identification, genetic screening, identification of pathological conditions, determination of patterns of specific gene expression, and others.

The present invention also provides methods for sequencing a segment of a polynucleotide comprising the steps of:

- a) combining:
- i) a substrate comprising a plurality of chemically synthesized and positionally distinguishable oligonucleotides capable of recognizing defined oligonucleotide sequences; and
- 25 ii) a target polynucleotide; thereby forming high fidelity matched duplex structures of complementary subsequences of known sequence; and
- 30 b) determining which of said reagents have specifically interacted with subsequences in said target polynucleotide.

35 In one embodiment, the segment is substantially the entire length of said polynucleotide.

In one embodiment, the substrates are beads. Preferably, the plurality of reagents comprise substantially

WO 92/10588

PCT/US91/09226

5

all possible subsequences of said preselected length found in said target. In another embodiment, the solid phase substrate is a single substrate having attached thereto reagents recognizing substantially all possible subsequences of preselected length found in said target.

In another embodiment, the method further comprises the step of analyzing a plurality of said recognized subsequences to assemble a sequence of said target polymer. In a bead embodiment, at least some of the plurality of substrates have one subsequence specific reagent attached thereto, and the substrates are coded to indicate the sequence specificity of said reagent.

The present invention also embraces a method of using a fluorescent nucleotide to detect interactions with oligonucleotide probes of known sequence, said method comprising:

- a) attaching said nucleotide to a target unknown polynucleotide sequence, and
- b) exposing said target polynucleotide sequence to a collection of positionally defined oligonucleotide probes of known sequences to determine the sequences of said probes which interact with said target.

In a further refinement, an additional step is included of:

- a) collating said known sequences to determine the overlaps of said known sequences to determine the sequence of said target sequence.

A method of mapping a plurality of sequences relative to one another is also provided, the method comprising:

- a) preparing a substrate having a plurality of positionally attached sequence specific probes are attached;
- b) exposing each of said sequences to said substrate, thereby determining the patterns of interaction between said sequence specific probes and said sequences; and

WO 92/10588

PCT/US91/09226

6

- c) determining the relative locations of said sequence specific probe interactions on said sequences to determine the overlaps and order of said sequences.

5 In one refinement, the sequence specific probes are oligonucleotides, applicable to where the target sequences are nucleic acid sequences.

In the nucleic acid sequencing application, the steps of the sequencing process comprise:

- 10 a) producing a matrix substrate having known positionally defined regions of known sequence specific oligonucleotide probes;
- 15 b) hybridizing a target polynucleotide to the positions on the matrix so that each of the positions which contain oligonucleotide probes complementary to a sequence on the target hybridize to the target molecule;
- 20 c) detecting which positions have bound the target, thereby determining sequences which are found on the target; and
- 25 d) analyzing the known sequences contained in the target to determine sequence overlaps and assembling the sequence of the target therefrom.

The enablement of the sequencing process by hybridization is based in large part upon the ability to synthesize a large number (e.g., to virtually saturate) of the possible overlapping sequence segments and distinguishing those probes which hybridize with fidelity from those which have mismatched bases, and to analyze a highly complex pattern of hybridization results to determine the overlap regions.

35 The detecting of the positions which bind the target sequence would typically be through a fluorescent label on the target. Although a fluorescent label is probably most convenient, other sorts of labels, e.g., radioactive, enzyme linked, optically detectable, or spectroscopic labels may be

WO 92/10588

PCT/US91/09226

7

used. Because the oligonucleotide probes are positionally defined, the location of the hybridized duplex will directly translate to the sequences which hybridize. Thus, upon analysis of the positions provides a collection of subsequences found within the target sequence. These subsequences are matched with respect to their overlaps so as to assemble an intact target sequence.

#### BRIEF DESCRIPTION OF THE FIGURES

Fig. 1 illustrates a flow chart for sequence, fingerprint, or mapping analysis.

Fig. 2 illustrates the proper function of a VLSIPS nucleotide synthesis.

Fig. 3 illustrates the proper function of a VLSIPS dinucleotide synthesis.

Fig. 4 illustrates the process of a VLSIPS trinucleotide synthesis.

#### DESCRIPTION OF THE PREFERRED EMBODIMENTS

- I. Overall Description
  - A. general
  - B. VLSIPS substrates
  - C. binary masking
  - D. applications
  - E. detection methods and apparatus
  - F. data analysis
- II. Theoretical Analysis
  - A. simple n-mer structure; theory
  - B. complications
- III. Polynucleotide Sequencing
  - A. preparation of substrate matrix
  - B. labeling target polynucleotide
  - C. hybridization conditions
  - D. detection; VLSIPS scanning
  - E. analysis
  - F. substrate reuse
- IV. Fingerprinting
  - A. general
  - B. preparation of substrate matrix
  - C. labeling target nucleotides
  - D. hybridization conditions
  - E. detection; VLSIPS scanning
  - F. analysis
  - G. substrate reuse
  - H. other polynucleotide aspects

SUBSTITUTE SHEET

WO 92/10588

PCT/US91/09226

8

- V. Mapping
- A. general
  - B. preparation of substrate matrix
  - C. labeling
  - D. hybridization/specific interaction
  - E. detection
  - F. analysis
  - G. substrate reuse
- VI. Additional Screening
- A. specific interactions
  - B. sequence comparisons
  - C. categorizations
  - D. statistical correlations
- VII. Formation of Substrate
- A. instrumentation
  - B. binary masking
  - C. synthetic methods
  - D. surface immobilization
- VIII. Hybridization/Specific Interaction
- A. general
  - B. important parameters
- IX. Detection Methods
- A. labeling techniques
  - B. scanning system
- X. Data Analysis
- A. general
  - B. hardware
  - C. software
- XI. Substrate Reuse
- A. removal of label
  - B. storage and preservation
  - C. processes to avoid degradation of oligomers
- XII. Integrated Sequencing Strategy
- A. initial mapping strategy
  - B. selection of smaller clones
- XIII. Commercial Applications
- A. sequencing
  - B. fingerprinting
  - C. mapping

\* \* \*

50

## I. OVERALL DESCRIPTION

A. General

The present invention relies in part on the ability to synthesize or attach specific recognition reagents at known

WO 92/10588

PCT/US91/09226

9

locations on a substrate, typically a single substrate. In particular, the present invention provides the ability to prepare a substrate having a very high density matrix pattern of positionally defined specific recognition reagents. The

5 reagents are capable of interacting with their specific targets while attached to the substrate, e.g., solid phase interactions, and by appropriate labeling of these targets, the sites of the interactions between the target and the specific reagents may be derived. Because the reagents are positionally

10 defined, the sites of the interactions will define the specificity of each interaction. As a result, a map of the patterns of interactions with specific reagents on the substrate is convertible into information on the specific interactions taking place, e.g., the recognized features.

15 Where the specific reagents recognize a large number of possible features, this system allows the determination of the combination of specific interactions which exist on the target molecule. Where the number of features is sufficiently large, the identical same combination, or pattern, of features is

20 sufficiently unlikely that a particular target molecule may often be uniquely defined by its features. In the extreme, the features may actually be the subunit sequence of the target molecule, and a given target sequence may be uniquely defined by its combination of features.

25 In particular, the methodology is applicable to sequencing polynucleotides. The specific sequence recognition reagents will typically be oligonucleotide probes which hybridize with specificity to subsequences found on the target sequence. A sufficiently large number of those probes allows

30 the fingerprinting of a target polynucleotide or the relative mapping of a collection of target polynucleotides, as described in greater detail below.

In the high resolution fingerprinting provided by a saturating collection of probes which include all possible

35 subsequences of a given size, e.g., 10-mers, collating of all the subsequences and determination of specific overlaps will be derived and the entire sequence can usually be reconstructed.

WO 92/10588

PCT/US91/09226

10

Sequence analysis may take the form of complete sequence determination, to the level of the sequence of individual subunits along the entire length of the target sequence. Sequence analysis also may take the form of sequence  
5 homology, e.g., less than absolute subunit resolution, where "similarity" in the sequence will be detectable, or the form of selective sequences of homology interspersed at specific or irregular locations.

In either case, the sequence is determinable at  
10 selective resolution or at particular locations. Thus, the hybridization method will be useful as a means for identification, e.g., a "fingerprint", much like a Southern hybridization method is used. It is also useful to map particular target sequences.

15

#### B. VLSIPS Substrates

The invention is enabled by the development of technology to prepare substrates on which specific reagents may be either positionally attached or synthesized. In particular,  
20 the very large scale immobilized polymer synthesis (VLSIPS) technology allows for the very high density production of an enormous diversity of reagents mapped out in a known matrix pattern on a substrate. These reagents specifically recognize subsequences in a target polymer and bind thereto, producing a  
25 map of positionally defined regions of interaction. These map positions are convertible into actual features recognized, and thus would be present in the target molecule of interest.

As indicated, the sequence specific recognition reagents will often be oligonucleotides which hybridize with  
30 fidelity and discrimination to the target sequence.

In the generic sense, the VLSIPS technology allows the production of a substrate with a high density matrix of positionally mapped regions with specific recognition reagents attached at each distinct region. By use of protective groups  
35 which can be positionally removed, or added, the regions can be activated or deactivated for addition of particular reagents or compounds. Details of the protection are described below and in PCT publication no. W090/15070, published December 13, 1990.

WO 92/10588

PCT/US91/09226

11

In a preferred embodiment, photosensitive protecting agents will be used and the regions of activation or deactivation may be controlled by electro-optical and optical methods, similar to many of the processes used in semiconductor wafer and chip fabrication.

In the nucleic acid nucleotide sequencing application, a VLSIPS substrate is synthesized having positionally defined oligonucleotide probes. See PCT publication no. WO90/15070, published December 13, 1990; and U.S.S.N. 07/624,120, filed December 6, 1990. By use of masking technology and photosensitive synthetic subunits, the VLSIPS apparatus allows for the stepwise synthesis of polymers according to a positionally defined matrix pattern. Each oligonucleotide probe will be synthesized at known and defined positional locations on the substrate. This forms a matrix pattern of known relationship between position and specificity of interaction. The VLSIPS technology allows the production of a very large number of different oligonucleotide probes to be simultaneously and automatically synthesized including numbers in excess of about  $10^2$ ,  $10^3$ ,  $10^4$ ,  $10^5$ ,  $10^6$ , or even more, and at densities of at least about  $10^2$ ,  $10^3/\text{cm}^2$ ,  $10^4/\text{cm}^2$ ,  $10^5/\text{cm}^2$  and up to  $10^6/\text{cm}^2$  or more. This application discloses methods for synthesizing polymers on a silicon or other suitably derivatized substrate, methods and chemistry for synthesizing specific types of biological polymers on those substrates, apparatus for scanning and detecting whether interaction has occurred at specific locations on the substrate, and various other technologies related to the use of a high density very large scale immobilized polymer substrate. In particular, sequencing, fingerprinting, and mapping applications are discussed herein in detail, though related technologies are described in U.S.S.N. 07/624,120, filed December 6, 1990; and PCT/US91/02989, filed May 1, 1991, each of which is hereby incorporated herein by reference.

The regions which define particular reagents will usually be generated by selective protecting groups which may be activated or deactivated. Typically the protecting group will be bound to a monomer subunit or spatial region, and can

WO 92/10588

PCT/US91/09226

12

be spatially affected by an activator, such as electromagnetic radiation. Examples of protective groups with utility herein include nitroveratryl oxycarbonyl (NVOC), nitrobenzyl oxycarbonyl (NBOC) or  $\alpha,\alpha$ -dimethyl-dimethoxybenzyl oxycarbonyl (DEZ).

### C. Binary Masking

In fact, the means for producing a substrate useful for these techniques are explained in U.S.S.N. 07/492,462 (VLSIPS CIP), which is hereby incorporated herein by reference. However, there are various particular ways to optimize the synthetic processes. Many of these methods are described in U.S.S.N. 07/624,120.

Briefly, the binary synthesis strategy refers to an ordered strategy for parallel synthesis of diverse polymer sequences by sequential addition of reagents which may be represented by a reactant matrix, and a switch matrix, the product of which is a product matrix. A reactant matrix is a  $1 \times n$  matrix of the building blocks to be added. The switch matrix is all or a subset of the binary numbers from 1 to  $n$  arranged in columns. In preferred embodiments, a binary strategy is one in which at least two successive steps illuminate half of a region of interest on the substrate. In most preferred embodiments, binary synthesis refers to a synthesis strategy which also factors a previous addition step. For example, a strategy in which a switch matrix for a masking strategy halves regions that were previously illuminated, illuminating about half of the previously illuminated region and protecting the remaining half (while also protecting about half of previously protected regions and illuminating about half of previously protected regions). It will be recognized that binary rounds may be interspersed with non-binary rounds and that only a portion of a substrate may be subjected to a binary scheme, but will still be considered to be a binary masking scheme within the definition herein. A binary "masking" strategy is a binary synthesis which uses light to remove protective groups from materials for addition of other materials such as nucleotides.

WO 92/10588

PCT/US91/09226

13

In particular, this procedure provides a simplified and highly efficient method for saturating all possible sequences of a defined length polymer. This masking strategy is also particularly useful in producing all possible  
5 oligonucleotide sequence probes of a given length.

#### D. Applications

The technology provided by the present invention has very broad applications. Although described specifically for  
10 polynucleotide sequences, similar sequencing, fingerprinting, mapping, and screening procedures may be applied to polypeptide, carbohydrate, or other polymers. This may be for de novo sequencing, or may be used in conjunction with a second sequencing procedure to provide independent verification. See,  
15 e.g., (1988) Science 242:1245. For example, a large polynucleotide sequence defined by either the Maxam and Gilbert technique or by the Sanger technique may be verified by using the present invention.

In addition, by selection of appropriate probes, a  
20 polynucleotide sequence can be fingerprinted. Fingerprinting is a less detailed sequence analysis which usually involves the characterization of a sequence by a combination of defined features. Sequence fingerprinting is particularly useful because the repertoire of possible features which can be tested  
25 is virtually infinite. Moreover, the stringency of matching is also variable depending upon the application. A Southern Blot analysis may be characterized as a means of simple fingerprint analysis.

Fingerprinting analysis may be performed to the  
30 resolution of specific nucleotides, or may be used to determine homologies, most commonly for large segments. In particular, an array of oligonucleotide probes of virtually any workable size may be positionally localized on a matrix and used to probe a sequence for either absolute complementary matching, or  
35 homology to the desired level of stringency using selected hybridization conditions.

In addition, the present invention provides means for mapping analysis of a target sequence or sequences. Mapping

WO 92/10588

PCT/US91/09226

14

will usually involve the sequential ordering of a plurality of various sequences, or may involve the localization of a particular sequence within a plurality of sequences. This may be achieved by immobilizing particular large segments onto the matrix and probing with a shorter sequence to determine which of the large sequences contain that smaller sequence. Alternatively, relatively shorter probes of known or random sequence may be immobilized to the matrix and a map of various different target sequences may be determined from overlaps. Principles of such an approach are described in some detail by Evans et al. (1989) "Physical Mapping of Complex Genomes by Cosmid Multiplex Analysis," Proc. Natl. Acad. Sci. USA 86:5030-5034; Michiels et al. (1987) "Molecular Approaches to Genome Analysis: A Strategy for the Construction of Ordered Overlap Clone Libraries," CABIOS 3:203-210; Olsen et al. (1986) "Random-Clone Strategy for Genomic Restriction Mapping in Yeast," Proc. Natl. Acad. Sci. USA 83:7826-7830; Craig, et al. (1990) "Ordering of Cosmid Clones Covering the Herpes Simplex Virus Type I (HSV-I) Genome: A Test Case for Fingerprinting by Hybridization," Nuc. Acids Res. 18:2653-2660; and Coulson, et al. (1986) "Toward a Physical Map of the Genome of the Nematode *Caenorhabditis elegans*," Proc. Natl. Acad. Sci. USA 83:7821-7825; each of which is hereby incorporated herein by reference.

Fingerprinting analysis also provides a means of identification. In addition to its value in apprehension of criminals from whom a biological sample, e.g., blood, has been collected, fingerprinting can ensure personal identification for other reasons. For example, it may be useful for identification of bodies in tragedies such as fire, flood, and vehicle crashes. In other cases the identification may be useful in identification of persons suffering from amnesia, or of missing persons. Other forensics applications include establishing the identity of a person, e.g., military identification "dog tags", or may be used in identifying the source of particular biological samples. Fingerprinting technology is described, e.g., in Carrano, et al. (1989) "A High-Resolution, Fluorescence-Based, Semi-automated method for DNA Fingerprinting," Genomics 4: 129-136, which is hereby

WO 92/10588

PCT/US91/09226

15

incorporated herein by reference. See, e.g., table I, for nucleic acid applications.

TABLE I

5 VLSIPS PROJECT IN NUCLEIC ACIDS

## I. Construction of Chips

## II. Applications

## A. Sequencing

- 1. Primary sequencing
- 10 2. Secondary sequencing (sequence checking)
- 3. Large scale mapping
- 4. Fingerprinting

## B. Duplex/Triplex formation

- 1. Antisense
- 15 2. Sequence specific function modulation  
(e.g. promoter inhibition)

## C. Diagnosis

- 1. Genetic markers
- 2. Type markers
- 20 a. Blood donors
- b. Tissue transplants

## D. Microbiology

- 1. Clinical microbiology
- 2. Food microbiology

25

## III. Instrumentation

## A. Chip machines

## B. Detection

## 30 IV. Software Development

- A. Instrumentation software
- B. Data reduction software
- C. Sequence analysis software

35 The fingerprinting analysis may be used to perform various types of genetic screening. For example, a single substrate may be generated with a plurality of screening probes, allowing for the simultaneous genetic screening for a

WO 92/10588

PCT/US91/09226

16

large number of genetic markers. Thus, prenatal or diagnostic screening can be simplified, economized, and made more generally accessible.

In addition to the sequencing, fingerprinting, and mapping applications, the present invention also provides means for determining specificity of interaction with particular sequences. Many of these applications were described in U.S.S.N. 07/362,901 (VLSIPS parent), U.S.S.N. 07/492,462 (VLSIPS CIP), U.S.S.N. 07/435,316 (caged biotin parent), and U.S.S.N. 07/612,671 (caged biotin CIP).

#### E. Detection Methods and Apparatus

An appropriate detection method applicable to the selected labeling method can be selected. Suitable labels include radionucleotides, enzymes, substrates, cofactors, inhibitors, magnetic particles, heavy metal atoms, and particularly fluorescers, chemilumescers, and spectroscopic labels. Patents teaching the use of such labels include U.S. Patent Nos. 3,817,837; 3,850,752; 3,939,350; 3,996,345; 4,277,437; 4,275,149; and 4,366,241.

With an appropriate label selected, the detection system best adapted for high resolution and high sensitivity detection may be selected. As indicated above, an optically detectable system, e.g., fluorescence or chemiluminescence would be preferred. Other detection systems may be adapted to the purpose, e.g., electron microscopy, scanning electron microscopy (SEM), scanning tunneling electron microscopy (STEM), infrared microscopy, atomic force microscopy (AFM), electrical conductance, and image plate transfer.

With a detection method selected, an apparatus for scanning the substrate will be designed. Apparatus, as described in PCT publication no. W090/15070, published December 13, 1990; or U.S.S.N. 07/624,120, filed December 6, 1990, are particularly appropriate. Design modifications may also be incorporated therein.

WO 92/10588

PCT/US91/09226

17

#### F. Data Analysis

Data is analyzed by processes similar to those described below in the section describing theoretical analysis. More efficient algorithms will be mathematically devised, and  
5 will usually be designed to be performed on a computer. Various computer programs which may more quickly or efficiently make measurement samples and distinguish signal from noise will also be devised. See, particularly, U.S.S.N. 07/624,120.

The initial data resulting from the detection system  
10 is an array of data indicative of fluorescent intensity versus location on the substrate. The data are typically taken over regions substantially smaller than the area in which synthesis of a given polymer has taken place. Merely by way of example, if polymers were synthesized in squares on the substrate having  
15 dimensions of 500 microns by 500 microns, the data may be taken over regions having dimensions of 5 microns by 5 microns. In most preferred embodiments, the regions over which fluorescence data are taken across the substrate are less than about 1/2 the area of the regions in which individual polymers are  
20 synthesized, preferably less than 1/10 the area in which a single polymer is synthesized, and most preferably less than 1/100 the area in which a single polymer is synthesized. Hence, within any area in which a given polymer has been synthesized, a large number of fluorescence data points are  
25 collected.

A plot of number of pixels versus intensity for a scan should bear a rough resemblance to a bell curve, but spurious data are observed, particularly at higher intensities. Since it is desirable to use an average of fluorescent  
30 intensity over a given synthesis region in determining relative binding affinity, these spurious data will tend to undesirably skew the data.

Accordingly, in one embodiment of the invention the data are corrected for removal of these spurious data points,  
35 and an average of the data points is thereafter utilized in determining relative binding efficiency. In general the data are fitted to a base curve and statistically measures are used to remove spurious data.

WO 92/10588

PCT/US91/09226

18

In an additional analytical tool, various degeneracy reducing analogues may be incorporated in the hybridization probes. Various aspects of this strategy are described, e.g., in Macevitz, S. (1990) PCT publication number WO 90/04652, which is hereby incorporated herein by reference.

## II. THEORETICAL ANALYSIS

The principle of the hybridization sequencing procedure is based, in part, upon the ability to determine overlaps of short segments. The VLSIPS technology provides the ability to generate reagents which will saturate the possible short subsequence recognition possibilities. The principle is most easily illustrated by using a binary sequence, such as a sequence of zeros and ones. Once having illustrated the application to a binary alphabet, the principle may easily be understood to encompass three letter, four letter, five or more letter, even 20 letter alphabets. A theoretical treatment of analysis of subsequence information to reconstruction of a target sequence is provided, e.e., in Lysov, Yu., et al. (1988) Doklady Akademi. Nauk. SSR 303:1508-1511; Khropko K., et al. (1989) FEBS Letters 256:118-122; Pevzner, P. (1989) J. of Biomolecular Structure and Dynamics 7:63-69; and Drmanac, R. et al. (1989) Genomics 4:114-128; each of which is hereby incorporated herein by reference.

The reagents for recognizing the subsequences will usually be specific for recognizing a particular polymer subsequence anywhere within a target polymer. It is preferable that conditions may be devised which allow absolute discrimination between high fidelity matching and very low levels of mismatching. The reagent interaction will preferably exhibit no sensitivity to flanking sequences, to the subsequence position within the target, or to any other remote structure within the sequence.

### A. Simple n-mer Structure: Theory

#### 1. Simple two letter alphabet: example

A simple example is presented below of how a sequence of ten digits comprising zeros and ones would be sequenceable

WO 92/10588

PCT/US91/09226

19

using short segments of five digits. For example, consider the sample ten digit sequence:

1010011100.

A VLSIPS substrate could be constructed, as discussed elsewhere, which would have reagents attached in a defined matrix pattern which specifically recognize each of the possible five digit sequences of ones and zeros. The number of possible five digit subsequences is  $2^5 = 32$ . The number of possible different sequences 10 digits long is  $2^{10} = 1,024$ . The five contiguous digit subsequences within a ten digit sequence number six, i.e., positioned at digits 1-5, 2-6, 3-7, 4-8, 5-9, and 6-10. It will be noted that the specific order of the digits in the sequence is important and that the order is directional, e.g., running left to right versus right to left. The first five digit sequence contained in the target sequence is 10100. The second is 01001, the third is 10011, the fourth is 00111, the fifth is 01110, and the sixth is 11100.

The VLSIPS substrate would have a matrix pattern of positionally attached reagents which recognize each of the different 5-mer subsequences. Those reagents which recognize each of the 6 contained 5-mers will bind the target, and a label allows the positional determination of where the sequence specific interaction has occurred. By correlation of the position in the matrix pattern, the corresponding bound subsequences can be determined.

In the above-mentioned sequence, six different 5-mer sequences would be determined to be present. They would be:

10100  
01001  
10011  
00111  
01110  
11100

Any sequence which contains the first five digit sequence, 10100, already narrows the number of possible sequences (e.g., from 1024 possible sequences) which contain it to less than about 192 possible sequences.

This 192 is derived from the observation that with the subsequence 10100 at the far left of the sequence, in positions 1-5, there are only 32 possible sequences. Likewise,

WO 92/10588

PCT/US91/09226

20

for that particular subsequence in positions 2-6, 3-7, 4-8, 5-9, and 6-10. So, to sum up all of the sequences that could contain 10100, there are 32 for each position and 6 positions for a total of about 192 possible sequences. However, some of these 10 digit sequences will have been counted twice. Thus, by virtue of containing the 10100 subsequence, the number of possible 10-mer sequences has been decreased from 1024 sequences to less than about 192 sequences.

In this example, not only do we know that sequence contains 10100, but we also know that it contains the second five character sequence, 01001. By virtue of knowing that the sequence contains 10100, we can look specifically to determine whether the sequence contains a subsequence of five characters which contains the four leftmost digits plus a next digit to the left. For example, we would look for a sequence of X1010, but we find that there is none. Thus, we know that the 10100 must be at the left end of the 10-mer. We would also look to see whether the sequence contains the rightmost four digits plus a next digit to the right, e.g., 0100X. We find that the sequence also contains the sequence 01001, and that X is a 1. Thus, we know at least that our target sequence has an overlap of 0100 and has the left terminal sequence 101001.

Applying the same procedure to the second 5-mer, we also know that the sequence must include a sequence of five digits having the sequence 1001Y where Y must be either 0 or 1. We look through the fragments and we see that we have a 10011 sequence within our target, thus Y is also 1. Thus, we would know that our sequence has a sequence of the first seven being 1010011.

Moving to the next 5-mer, we know that there must be a sequence of 0011Z, where Z must be either 0 or 1. We look at the fragments produced above and see that the target sequence contains a 00111 subsequence and Z is 1. Thus, we know the sequence must start with 10100111.

The next 5-mer must be of the sequence 0111W where W must be 0 or 1. Again, looking up at the fragments produced, we see that the target sequence contains a 01110 subsequence, and W is a 0. Thus, our sequence to this point is 101001110.

WO 92/10588

PCT/US91/09226

21

We know that the last 5-mer must be either 11100 or 11101. Looking above, we see that it is 11100 and that must be the last of our sequence. Thus, we have determined that our sequence must have been 1010011100.

5                However, it will be recognized from the example above with the sequences provided therein, that the sequence analysis can start with any known positive probe subsequence. The determination may be performed by moving linearly along the sequence checking the known sequence with a limited number of  
10 next positions. Given this possibility, the sequence may be determined, besides by scanning all possible oligonucleotide probe positions, by specifically looking only where the next possible positions would be. This may increase the complexity of the scanning but may provide a longer time span dedicated  
15 towards scanning and detecting specific positions of interest relative to other sequence possibilities. Thus, the scanning apparatus could be set up to work its way along a sequence from a given contained oligonucleotide to only look at those positions on the substrate which are expected to have a  
20 positive signal.

              It is seen that given a sequence, it can be deconstructed into n-mers to produce a set of internal contiguous subsequences. From any given target sequence, we would be able to determine what fragments would result. The hybridization  
25 sequence method depends, in part, upon being able to work in the reverse, from a set of fragments of known sequences to the full sequence. In simple cases, one is able to start at a single position and work in either or both directions towards the ends of the sequence as illustrated in the example.

30                The number of possible sequences of a given length increases very quickly with the length of that sequence. Thus, a 10-mer of zeros and ones has 1024 possibilities, a 12-mer has 4096. A 20-mer has over a million possibilities, and a 30-mer has over a billion. However, a given 30-mer has, at most, 26  
35 different internal 5-mer sequences. Thus, a 30 character target sequence having over a million possible sequences can be substantially defined by only 26 different 5-mers. It will be recognized that the probe oligonucleotides will preferably, but

WO 92/10588

PCT/US91/09226

22

need not necessarily, be of identical length, and that the probe sequences need not necessarily be contiguous in that the overlapping subsequences need not differ by only a single subunit. Moreover, each position of the matrix pattern need  
5 not be homogeneous, but may actually contain a plurality of probes of known sequence. In addition, although all of the possible subsequence specifications would be preferred, a less than full set of sequences specifications could be used. In particular, although a substantial fraction will preferably be  
10 at least about 70%, it may be less than that. About 20% would be preferred, more preferably at least about 30% would be desired. Higher percentages would be especially preferred.

## 2. Example of four letter alphabet

15 A four letter alphabet may be conceptualized in at least two different ways from the two letter alphabet. One way, is to consider the four possible values at each position and to analogize in a similar fashion to the binary example each of the overlaps. A second way is to group the binary  
20 digits into groups.

Using the first means, the overlap comparisons are performed with a four letter alphabet rather than a two letter alphabet. Then, in contrast to the binary system with 10 positions where  $2^{10} = 1024$  possible sequences, in a 4-character  
25 alphabet with 10 positions, there will actually be  $4^{10} = 1,048,576$  possible sequences. Thus, the complexity of a four character sequence has a much larger number of possible sequences compared to a two character sequence. Note, however, that there are still only 6 different internal 5-mers. For  
30 simplicity, we shall examine a 5 character string with 3 character subsequences. Instead of only 1 and 0, the characters may be designated, e.g., A, C, G, and T. Let us take the sequence GGCTA. The 3-mer subsequences are:

35 GGC  
GCT  
CTA

Given these subsequences, there is one sequence, or at most only a few sequences which would produce that combination of  
40 subsequences, i.e., GGCTA.

WO 92/10588

PCT/US91/09226

23

Alternatively, with a four character universe, the binary system can be looked at in pairs of digits. The pairs would be 00, 01, 10, and 11. In this manner, the earlier used sequence 1010011100 is looked at as 10,10,01,11,00. Then the first character of two digits is selected from the possible universe of the four representations 00, 01, 10, and 11. Then a probe would be in an even number of digits, e.g., not five digits, but, three pairs of digits or six digits. A similar comparison is performed and the possible overlaps determined.

10 The 3-pair subsequences are:

10,10,01  
10,01,11  
01,11,00

and the overlap reconstruction produces 10,10,01,11,00.

15 The latter of the two conceptual views of the 4 letter alphabet provides a representation which is similar to what would be provided in a digital computer. The applicability to a four nucleotide alphabet is easily seen by assigning, e.g., 00 to A, 01 to C, 10 to G, and 11 to T. And, in fact, if such a correspondence is used, both examples for the 4 character sequences can be seen to represent the same target sequence. The applicability of the hybridization method and its analysis for determining the ultimate sequence is easily seen if A is the representation of adenine, C is the representation of cytosine, G is the representation of guanine, and T is the representation of thymine or uracil.

#### B. Complications

Two obvious complications exist with the method of sequence analysis by hybridization. The first results from a probe of inappropriate length while the second relates to internally repeated sequences.

The first obvious complication is a problem which arises from an inappropriate length of recognition sequence, which causes problems with the specificity of recognition. For example, if the recognized sequence is too short, every sequence which is utilized will be recognized by every probe sequence. This occurs, e.g., in a binary system where the probes are each of sequences which occur relatively frequently,

WO 92/10588

PCT/US91/09226

24

e.g., a two character probe for the binary system. Each possible two character probe would be expected to appear  $\frac{1}{4}$  of the time in every single two character position. Thus, the above sequence example would be recognized by each of the 00, 10, 01, and 11. Thus, the sequence information is virtually lost because the resolution is too low and each recognition reagent specifically binds at multiple sites on the target sequence.

The number of different probes which bind to a target depends on the relationship between the probe length and the target length. At the extreme of short probe length, the just mentioned problem exists of excessive redundancy and lack of resolution. The lack of stability in recognition will also be a problem with extremely short probes. At the extreme of long probe length, each entire probe sequence is on a different position of a substrate. However, a problem arises from the number of possible sequences, which goes up dramatically with the length of the sequence. Also, the specificity of recognition begins to decrease as the contribution to binding by any particular subunit may become sufficiently low that the system fails to distinguish the fidelity of recognition. Mismatched hybridization may be a problem with the polynucleotide sequencing applications, though the fingerprinting and mapping applications may not be so strict in their fidelity requirements. As indicated above, a thirty position binary sequence has over a million possible sequences, a number which starts to become unreasonably large in its required number of different sequences, even though the target length is still very short. Preparing a substrate with all sequence possibilities for a long target may be extremely difficult due to the many different oligomers which must be synthesized.

The above example illustrates how a long target sequence may be reconstructed with a reasonably small number of shorter subsequences. Since the present day resolution of the regions of the substrate having defined oligomer probes attached to the substrate approaches about 10 microns by 10 microns for resolvable regions, about  $10^6$ , or 1 million,

WO 92/10588

PCT/US91/09226

25

positions can be placed on a one centimeter square substrate. However, high resolution systems may have particular disadvantages which may be outweighed using the lower density substrate matrix pattern. For this reason, a sufficiently

5 large number of probe sequences can be utilized so that any given target sequence may be determined by hybridization to a relatively small number of probes.

A second complication relates to convergence of sequences to a single subsequence. This will occur when a

10 particular subsequence is repeated in the target sequence. This problem can be addressed in at least two different ways. The first, and simpler way, is to separate the repeat sequences onto two different targets. Thus, each single target will not have the repeated sequence and can be analyzed to its end.

15 This solution, however, complicates the analysis by requiring that some means for cutting at a site between the repeats can be located. Typically a careful sequencer would want to have two intermediate cut points so that the intermediate region can also be sequenced in both directions across each of the cut

20 points. This problem is inherent in the hybridization method for sequencing but can be minimized by using a longer known probe sequence so that the frequency of probe repeats is decreased.

Knowing the sequence of flanking sequences of the

25 repeat will simplify the use of polymerase chain reaction (PCR) or a similar technique to further definitively determine the sequence between sequence repeats. Probes can be made to hybridize to those known sequences adjacent the repeat sequences, thereby producing new target sequences for analysis.

30 See, e.g., Innis et al. (eds.) (1990) PCR Protocols: A Guide to Methods and Applications, Academic Press; and methods for synthesis of oligonucleotide probes, see, e.g., Gait (1984) Oligonucleotide Synthesis: A Practical Approach, IRL Press, Oxford.

35 Other means for dealing with convergence problems include using particular longer probes, and using degeneracy reducing analogues, see, e.g., Macevicz, S. (1990) PCT publication number WO 90/04652, which is hereby incorporated

WO 92/10588

PCT/US91/09226

26

herein by reference. By use of stretches of the degeneracy reducing analogues with other probes in particular combinations, the number of probes necessary to fully saturate the possible oligomer probes is decreased. For example, with a stretch of 12-mers having the central 4-mer of degenerate nucleotides, in combination with all of the possible 8-mers, the collection numbers twice the number of possible 8-mers, e.g.  $65,536 + 65,536 = 131,072$ , but the population provides screening equivalent to all possible 12-mers.

By way of further explanation, all possible oligonucleotide 8-mers may be depicted in the fashion:

N1-N2-N3-N4-N5-N6-N7-N8,

in which there are  $4^8 = 65,536$  possible 8-mers. As described in U.S.S.N. 07/624,120, producing all possible 8-mers requires  $4 \times 8 = 32$  chemical binary synthesis steps to produce the entire matrix pattern of 65,536 8-mer possibilities. By incorporating degeneracy reducing nucleotides, D's, which hybridize nonselectively to any corresponding complementary nucleotide, new oligonucleotides 12-mers can be made in the fashion:

N1-N2-N3-N4-D-D-D-D-N5-N6-N7-N8,

in which there are again, as above, only  $4^8 = 65,536$  possible "12-mers", which in reality only have 8 different nucleotides.

However, it can be seen that each possible 12-mer probe could be represented by a group of the two 8-mer types. Moreover, repeats of less than 12 nucleotides would not converge, or cause repeat problems in the analysis. Thus, instead of requiring a collection of probes corresponding to all 12-mers, or  $4^{12} = 16,777,216$  different 12-mers, the same information can be derived by making 2 sets of "8-mers" consisting of the typical 8-mer collection of  $4^8 = 65,536$  and the "12-mer" set with the degeneracy reducing analogues, also requiring making  $4^8 = 65,536$ . The combination of the two sets, requires making  $65,536 + 65,536 = 131,072$  different molecules, but giving the information of 16,777,216 molecules. Thus, incorporating the degeneracy reducing analogue decreases the number of molecules necessary to get 12-mer resolution by a factor of about 128-fold.

WO 92/10588

PCT/US91/09226

27

### III. POLYNUCLEOTIDE SEQUENCING

In principle, the making of a substrate having a positionally defined matrix pattern of all possible oligonucleotides of a given length involves a conceptually simple method of synthesizing each and every different possible oligonucleotide, and affixed to a definable position. Oligonucleotide synthesis is presently mechanized and enabled by current technology, see, e.g., U.S.S.N. 07/362,901 (VLSIPS parent); U.S.S.N. 07/492,462 (VLSIPS CIP); and instruments supplied by Applied Biosystems, Foster City, California.

#### A. Preparation of Substrate Matrix

The production of the collection of specific oligonucleotides used in polynucleotide sequencing may be produced in at least two different ways. Present technology certainly allows production of ten nucleotide oligomers on a solid phase or other synthesizing system. See, e.g., instrumentation provided by Applied Biosystems, Foster City, California. Although a single oligonucleotide can be relatively easily made, a large collection of them would typically require a fairly large amount of time and investment. For example, there are  $4^{10} = 1,048,576$  possible ten nucleotide oligomers. Present technology allows making each and every one of them in a separate purified form though such might be costly and laborious.

Once the desired repertoire of possible oligomer sequences of a given length have been synthesized, this collection of reagents may be individually positionally attached to a substrate, thereby allowing a batchwise hybridization step. Present technology also would allow the possibility of attaching each and every one of these 10-mers to a separate specific position on a solid matrix. This attachment could be automated in any of a number of ways, particularly use of a caged biotin type linking. This would produce a matrix having each of different possible 10-mers.

A batchwise hybridization is much preferred because of its reproducibility and simplicity. An automated process of

WO 92/10588

PCT/US91/09226

28

attaching various reagents to positionally defined sites on a substrate is provided in PCT publication no. W090/15070; U.S.S.N. 07/624,120; and PCT publication no. W091/07087; each of which is hereby incorporated herein by reference.

5           Instead of separate synthesis of each oligonucleotide, these oligonucleotides are conveniently synthesized in parallel by sequential synthetic processes on a defined matrix pattern as provided in PCT publication no. W090/15070; and U.S.S.N. 07/624,120, which are incorporated  
10           herein by reference. Here, the oligonucleotides are synthesized stepwise on a substrate at positionally separate and defined positions. Use of photosensitive blocking reagents allows for defined sequences of synthetic steps over the surface of a matrix pattern. By use of the binary masking  
15           strategy, the surface of the substrate can be positioned to generate a desired pattern of regions, each having a defined sequence oligonucleotide synthesized and immobilized thereto.

          Although the prior art technology can be used to generate the desired repertoire of oligonucleotide probes, an  
20           efficient and cost effective means would be to use the VLSIPS technology described in PCT publication no. W090/15070 and U.S.S.N. 07/624,120. In this embodiment, the photosensitive reagents involved in the production of such a matrix are described below.

25           The regions for synthesis may be very small, usually less than about 100  $\mu\text{m}$  x 100  $\mu\text{m}$ , more usually less than about 50  $\mu\text{m}$  x 50  $\mu\text{m}$ . The photolithography technology allows synthetic regions of less than about 10  $\mu\text{m}$  x 10  $\mu\text{m}$ , about 3  $\mu\text{m}$  x 3  $\mu\text{m}$ , or less. The detection also may detect such sized  
30           regions, though larger areas are more easily and reliably measured.

          At a size of about 30 microns by 30 microns, one million regions would take about 11 centimeters square or a single wafer of about 4 centimeters by 4 centimeters. Thus the  
35           present technology provides for making a single matrix of that size having all one million plus possible oligonucleotides. Region size are sufficiently small to correspond to densities of at least about 5 regions/cm<sup>2</sup>, 20 regions/cm<sup>2</sup>, 50 regions/cm<sup>2</sup>,

WO 92/10588

PCT/US91/09226

29

100 regions/cm<sup>2</sup>, and greater, including 300 regions/cm<sup>2</sup>, 1000 regions/cm<sup>2</sup>, 3K regions/cm<sup>2</sup>, 10K regions/cm<sup>2</sup>, 30K regions/cm<sup>2</sup>, 100K regions/cm<sup>2</sup>, 300K regions/cm<sup>2</sup> or more, even in excess of one million regions/cm<sup>2</sup>.

5           Although the pattern of the regions which contain specific sequences is theoretically not important, for practical reasons certain patterns will be preferred in synthesizing the oligonucleotides. The application of binary masking algorithms for generating the pattern of known  
10 oligonucleotide probes is described in related U.S.S.N. 07/624,120. By use of these binary masks, a highly efficient means is provided for producing the substrate with the desired matrix pattern of different sequences. Although the binary masking strategy allows for the synthesis of all lengths of  
15 polymers, the strategy may be easily modified to provide only polymers of a given length. This is achieved by omitting steps where a subunit is not attached.

          The strategy for generating a specific pattern may take any of a number of different approaches. These approaches  
20 are well described in related application U.S.S.N. 07/624,120 and include a number of binary masking approaches which will not be exhaustively discussed herein. However, the binary masking and binary synthesis approaches provide a maximum of diversity with a minimum number of actual synthetic steps.

25           The length of oligonucleotides used in sequencing applications will be selected on criteria determined to some extent by the practical limits discussed above. For example, if probes are made as oligonucleotides, there will be 65,536 possible eight nucleotide sequences. If a nine subunit  
30 oligonucleotide is selected, there are 262,144 possible permutations of sequences. If a ten-mer oligonucleotide is selected, there are 1,048,576 possible permutations of sequences. As the number gets larger, the required number of positionally defined subunits necessary to saturate the  
35 possibilities also increases. With respect to hybridization conditions, the length of the matching necessary to converse stability of the conditions selected can be compensated for.

WO 92/10588

PCT/US91/09226

30

See, e.g., Kanehisa, M. (1984) Nuc. Acids Res. 12:203-213, which is hereby incorporated herein by reference.

Although not described in detail here, but below for oligonucleotide probes, the VLSIPS technology would typically use a photosensitive protective group on an oligonucleotide. Sample oligonucleotides are shown in Figure 1. In particular, the photoprotective group on the nucleotide molecules may be selected from a wide variety of positive light reactive groups preferably including nitro aromatic compounds such as o-nitrobenzyl derivatives or benzyisulfonyl. See, e.g., Gait (1984) Oligonucleotide Synthesis: A Practical Approach, IRL Press, Oxford, which is hereby incorporated herein by reference. In a preferred embodiment, 6-nitro-veratryl oxycarbony (NVOC), 2-nitrobenzyl oxycarbonyl (NBOC), or  $\alpha,\alpha$ -dimethyl-dimethoxybenzyl oxycarbonyl (DEZ) is used. Photoremovable protective groups are described in, e.g., Patchornik (1970) J. Amer. Chem. Soc. 92:6333; and Amit et al. (1974) J. Organic Chem. 39:192; each of which is hereby incorporated herein by reference.

A preferred linker for attaching the oligonucleotide to a silicon matrix is illustrated in Figure 2. A more detailed description is provided below. A photosensitive blocked nucleotide may be attached to specific locations of unblocked prior cycles of attachments on the substrate and can be successively built up to the correct length oligonucleotide probe.

It should be noted that multiple substrates may be simultaneously exposed to a single target sequence where each substrate is a duplicate of one another or where, in combination, multiple substrates together provide the complete or desired subset of possible subsequences. This provides the opportunity to overcome a limitation of the density of positions on a single substrate by using multiple substrates. In the extreme case, each probe might be attached to a single bead or substrate and the beads sorted by whether there is a binding interaction. Those beads which do bind might be encoded to indicate the subsequence specificity of reagents attached thereto.

WO 92/10588

PCT/US91/09226

31

Then, the target may be bound to the whole collection of beads and those beads that have appropriate specific reagents on them will bind to target. Then a sorting system may be utilized to sort those beads that actually bind the target from those that do not. This may be accomplished by presently available cell sorting devices or a similar apparatus. After the relatively small number of beads which have bound the target have been collected, the encoding scheme may be read off to determine the specificity of the reagent on the bead. An encoding system may include a magnetic system, a shape encoding system, a color encoding system, or a combination of any of these, or any other encoding system. Once again, with the collection of specific interactions that have occurred, the binding may be analyzed for sequence information, fingerprint information, or mapping information.

The parameters of polynucleotide sizes of both the probes and target sequences are determined by the applications and other circumstances. The length of the oligonucleotide probes used will depend in part upon the limitations of the VLSIPS technology to provide the number of desired probes. For example, in an absolute sequencing application, it is often useful to have virtually all of the possible oligonucleotides of a given length. As indicated above, there are 65,536 8-mers, 262,144 9-mers, 1,048,576 10-mers, 4,194,304 11-mers, etc. As the length of the oligomer increases the number of different probes which must be synthesized also increases at a rate of a factor of 4 for every additional nucleotide. Eventually the size of the matrix and the limitations in the resolution of regions in the matrix will reach the point where an increase in number of probes becomes disadvantageous. However, this sequencing procedure requires that the system be able to distinguish, by appropriate selection of hybridization and washing conditions, between binding of absolute fidelity and binding of complementary sequences containing mismatches. On the other hand, if the fidelity is unnecessary, this discrimination is also unnecessary and a significantly longer probe may be used. Significantly longer probes would typically be useful in fingerprinting or mapping applications.

WO 92/10588

PCT/US91/09226

32

The length of the probe is selected for a length that it will bind with specificity to possible targets. The hybridization conditions are also very important in that they will determine how close the homology of complementary binding will be detected. In fact, a single target may be evaluated at a number of different conditions to determine its spectrum of specificity for binding particular probes. This may find use in a number of other applications besides the polynucleotide sequencing fingerprinting or mapping. In a related fashion, different regions with reagents having differing affinities or levels of specificity may allow such a spectrum to be defined using a single incubation, where various regions, at a given hybridization condition, show the binding affinity. For example, fingerprint probes of various lengths, or with specific defined non-matches may be used. Unnatural nucleotides or nucleotides exhibiting modified specificity of complementary binding are described in greater detail in Macevicz (1990) PCT pub. No. WO 90/04652; and see the section on modified nucleotides in the Sigma Chemical Company catalogue.

#### B. Labeling Target Nucleotide

The label used to detect the target sequences will be determined, in part, by the detection methods being applied. Thus, the labeling method and label used are selected in combination with the actual detecting systems being used.

Once a particular label has been selected, appropriate labeling protocols will be applied, as described below for specific embodiments. Standard labeling protocols for nucleic acids are described, e.g., in Sambrook et al.; Kambara, H. et al. (1988) BioTechnology 6:816-821; Smith, L. et al. (1985) Nuc. Acids Res. 13:2399-2412; for polypeptides, see, e.g., Allen G. (1989) Sequencing of Proteins and Peptides, Elsevier, New York, especially chapter 5, and Greenstein and Winitz (1961) Chemistry of the Amino Acids, Wiley and Sons, New York. Carbohydrate labeling is described, e.g., in Chaplin and Kennedy (1986) Carbohydrate Analysis: A Practical Approach, IRL Press, Oxford. Labeling of other polymers will be

WO 92/10588

PCT/US91/09226

33

performed by methods applicable to them as recognized by a person having ordinary skill in manipulating the corresponding polymer.

In some embodiments, the target need not actually be labeled if a means for detecting where interaction takes place is available. As described below, for a nucleic acid embodiment, such may be provided by an intercalating dye which intercalates only into double stranded segments, e.g., where interaction occurs. See, e.g., Sheldon et al. U.S. Pat. No. 4,582,789.

In many uses, the target sequence will be absolutely homogeneous, both with respect to the total sequence and with respect to the ends of each molecule. Homogeneity with respect to sequence is important to avoid ambiguity. It is preferable that the target sequences of interest not be contaminated with a significant amount of labeled contaminating sequences. The extent of allowable contamination will depend on the sensitivity of the detection system and the inherent signal to noise of the system. Homogeneous contamination sequences will be particularly disruptive of the sequencing procedure.

However, although the target polynucleotide must have a unique sequence, the target molecules need not have identical ends. In fact, the homogeneous target molecule preparation may be randomly sheared to increase the numerical number of molecules. Since the total information content remains the same, the shearing results only in a higher number of distinct sequences which may be labeled and bind to the probe. This fragmentation may give a vastly superior signal relative to a preparation of the target molecules having homogeneous ends. The signal for the hybridization is likely to be dependent on the numerical frequency of the target-probe interactions. If a sequence is individually found on a larger number of separate molecules a better signal will result. In fact, shearing a homogeneous preparation of the target may often be preferred before the labeling procedure is performed, thereby producing a large number of labeling groups associated with each subsequence.

WO 92/10588

PCT/US91/09226

34

### C. Hybridization Conditions

The hybridization conditions between probe and target should be selected such that the specific recognition interaction, i.e., hybridization, of the two molecules is both sufficiently specific and sufficiently stable. See, e.g., Hames and Higgins (1985) Nucleic Acid Hybridisation: A Practical Approach, IRL Press, Oxford. These conditions will be dependent both on the specific sequence and often on the guanine and cytosine (GC) content of the complementary hybrid strands. The conditions may often be selected to be universally equally stable independent of the specific sequences involved. This typically will make use of a reagent such as an arylammonium buffer. See, Wood et al. (1985) "Base Composition-independent Hybridization in Tetramethylammonium Chloride: A Method for Oligonucleotide Screening of Highly Complex Gene Libraries," Proc. Natl. Acad. Sci. USA, 82:1585-1588; and Krupov et al. (1989) "An Oligonucleotide Hybridization Approach to DNA Sequencing," FEBS Letters, 256:118-122; each of which is hereby incorporated herein by reference. An arylammonium buffer tends to minimize differences in hybridization rate and stability due to GC content. By virtue of the fact that sequences then hybridize with approximately equal affinity and stability, there is relatively little bias in strength or kinetics of binding for particular sequences. Temperature and salt conditions along with other buffer parameters should be selected such that the kinetics of renaturation should be essentially independent of the specific target subsequence or oligonucleotide probe involved. In order to ensure this, the hybridization reactions will usually be performed in a single incubation of all the substrate matrices together exposed to the identical same target probe solution under the same conditions.

Alternatively, various substrates may be individually treated differently. Different substrates may be produced, each having reagents which bind to target subsequences with substantially identical stabilities and kinetics of hybridization. For example, all of the high GC content probes could be synthesized on a single substrate which is treated

WO 92/10588

PCT/US91/09226

35

accordingly. In this embodiment, the arylammonium buffers could be unnecessary. Each substrate is then treated in a manner that the collection of substrates show essentially uniform binding and the hybridization data of target binding to the individual substrate matrix is combined with the data from other substrates to derive the necessary subsequence binding information. The hybridization conditions will usually be selected to be sufficiently specific that the fidelity of base matching will be properly discriminated. Of course, control hybridizations should be included to determine the stringency and kinetics of hybridization.

#### D. Detection: VLSIPS Scanning

The next step of the sequencing process by hybridization involves labeling of target polynucleotide molecules. A quickly and easily detectable signal is preferred. The VLSIPS apparatus is designed to easily detect a fluorescent label, so fluorescent tagging of the target sequence is preferred. Other suitable labels include heavy metal labels, magnetic probes, chromogenic labels (e.g., phosphorescent labels, dyes, and fluorophores) spectroscopic labels, enzyme linked labels, radioactive labels, and labeled binding proteins. Additional labels are described in U.S. Pat. No. 4,366,241, which is incorporated herein by reference.

The detection methods used to determine where hybridization has taken place will typically depend upon the label selected above. Thus, for a fluorescent label a fluorescent detection step will typically be used. PCT publication no. WO90/15070 and U.S.S.N. 07/624,120 describe apparatus and mechanisms for scanning a substrate matrix using fluorescence detection, but a similar apparatus is adaptable for other optically detectable labels.

The detection method provides a positional localization of the region where hybridization has taken place. However, the position is correlated with the specific sequence of the probe since the probe has specifically been attached or synthesized at a defined substrate matrix position. Having collected all of the data indicating the subsequences present

WO 92/10588

PCT/US91/09226

36

in the target sequence, this data may be aligned by overlap to reconstruct the entire sequence of the target, as illustrated above.

It is also possible to dispense with actual labeling if some means for detecting the positions of interaction between the sequence specific reagent and the target molecule are available. This may take the form of an additional reagent which can indicate the sites either of interaction, or the sites of lack of interaction, e.g., a negative label. For the nucleic acid embodiments, locations of double strand interaction may be detected by the incorporation of intercalating dyes, or other reagents such as antibody or other reagents that recognize helix formation, see, e.g., Sheldon, et al. (1986) U.S. Pat. No. 4,582,789, which is hereby incorporated herein by reference.

#### E. Analysis

Although the reconstruction can be performed manually as illustrated above, a computer program will typically be used to perform the overlap analysis. A program may be written and run on any of a large number of different computer hardware systems. The variety of operating systems and languages useable will be recognized by a computer software engineer. Various different languages may be used, e.g., BASIC; C; PASCAL; etc. A simple flow chart of data analysis is illustrated in Figure 4.

#### F. Substrate Reuse

Finally, after a particular sequence has been hybridized and the pattern of hybridization analyzed, the matrix substrate should be reusable and readily prepared for exposure to a second or subsequent target polynucleotides. In order to do so, the hybrid duplexes are disrupted and the matrix treated in a way which removes all traces of the original target. The matrix may be treated with various detergents or solvents to which the substrate, the oligonucleotide probes, and the linkages to the substrate are inert. This treatment may include an elevated temperature

WO 92/10588

PCT/US91/09226

37

treatment, treatment with organic or inorganic solvents, modifications in pH, and other means for disrupting specific interaction. Thereafter, a second target may actually be applied to the recycled matrix and analyzed as before.

5

#### IV. FINGERPRINTING

##### A. General

Many of the procedures and techniques used in the polynucleotide sequencing section are also appropriate for fingerprinting applications. See, e.g., Poustka, et al. (1986) Cold Spring Harbor Symposia on Quant. Biol., vol. LI, 131-139, Cold Spring Harbor Press, New York; which is hereby incorporated herein by reference. The fingerprinting method provided herein is based, in part, upon the ability to positionally localize a large number of different specific probes onto a single substrate. This high density matrix pattern provides the ability to screen for, or detect, a very large number of different sequences simultaneously. In fact, depending upon the hybridization conditions, fingerprinting to the resolution of virtually absolute matching of sequence is possible thereby approaching an absolute sequencing embodiment. And the sequencing embodiment is very useful in identifying the probes useful in further fingerprinting uses. For example, characteristic features of genetic sequences will be identified as being diagnostic of the entire sequence. However, in most embodiments, longer probe and target will be used, and for which slight mismatching may not need to be resolved.

25

##### B. Preparation of Substrate Matrix

A collection of specific probes may be produced by either of the methods described above in the section on sequencing. Specific oligonucleotide probes of desired lengths may be individually synthesized on a standard oligonucleotide synthesizer. The length of these probes is limited only by the length of the ability of the synthesizer to continue to accurately synthesize a molecule. Oligonucleotides or sequence fragments may also be isolated from natural sources. Biological amplification methods may be coupled with synthetic

35

30

WO 92/10588

PCT/US91/09226

38

synthesizing procedures such as, e.g., polymerase chain reaction.

In one embodiment, the individually isolated probes may be attached to the matrix at defined positions. These probe reagents may be attached by an automated process making use of the caged biotin methodology described in U.S.S.N. 07/612,671 (caged biotin CIP), or using photochemical reagents, see, e.g., Dattagupta et al. (1985) U.S. Pat. No. 4,542,102 and (1987) U.S. Pat. No. 4,713,326. Each individual purified reagent can be attached individually at specific locations on a substrate.

In another embodiment, the VLSIPS synthesizing technique may be used to synthesize the desired probes at specific positions on a substrate. The probes may be synthesized by successively adding appropriate monomer subunits, e.g., nucleotides, to generate the desired sequences.

In another embodiment, a relatively short specific oligonucleotide is used which serves as a targeting reagent for positionally directing the sequence recognition reagent. For example, the sequence specific reagents having a separate additional sequence recognition segment (usually of a different polymer from the target sequence) can be directed to target oligonucleotides attached to the substrate. By use of non-natural targeting reagents, e.g., unusual nucleotide analogues which pair with other unnatural nucleotide analogues and which do not interfere with natural nucleotide interactions, the natural and non-natural portions can coexist on the same molecule without interfering with their individual functionalities. This can combine both a synthetic and biological production system analogous to the technique for targeting monoclonal antibodies to locations on a VLSIPS substrate at defined positions. Unnatural optical isomers of nucleotides may be useful unnatural reagents subject to similar chemistry, but incapable of interfering with the natural biological polymers. See also, U.S.S.N. 07/626,730, filed December 6, 1990; which is hereby incorporated herein by reference.

WO 92/10588

PCT/US91/09226

39

After the separate substrate attached reagents are attached to the targeting segment, the two are crosslinked, thereby permanently attaching them to the substrate. Suitable crosslinking reagents are known, see, e.g., Dattagupta et al. (1985) U.S. Pat. No. 4,542,102 and (1987) "Coupling of nucleic acids to solid support by photochemical methods," U.S. Pat. No. 4,713,326, each of which is hereby incorporated herein by reference. Similar linkages for attachment of proteins to a solid substrate are provided, e.g., in Merrifield (1986) Science 232:341, which is hereby incorporated herein by reference.

#### C. Labeling Target Nucleotides

The labeling procedures used in the sequencing embodiments will also be applicable in the fingerprinting embodiments. However, since the fingerprinting embodiments often will involve relatively large target molecules and relatively short oligonucleotide probes, the amount of signal necessary to incorporate into the target sequence may be less critical than in the sequencing applications. For example, a relatively long target with a relatively small number of labels per molecule may be easily amplified or detected because of the relatively large target molecule size.

In various embodiments, it may be desired to cleave the target into smaller segments as in the sequencing embodiments. The labeling procedures and cleavage techniques described in the sequencing embodiments would usually also be applicable here.

#### D. Hybridization Conditions

The hybridization conditions used in fingerprinting embodiments will typically be less critical than for the sequencing embodiments. The reason is that the amount of mismatching which may be useful in providing the fingerprinting information would typically be far greater than that necessary in sequencing uses. For example, Southern hybridizations do not typically distinguish between slightly mismatched sequences. Under these circumstances, important and valuable

WO 92/10588

PCT/US91/09226

40

information may be arrived at with less stringent hybridization conditions while providing valuable fingerprinting information. However, since the entire substrate is typically exposed to the target molecule at one time, the binding affinity of the probes should usually be of approximately comparable levels. For this reason, if oligonucleotide probes are being used, their lengths should be approximately comparable and will be selected to hybridize under conditions which are common for most of the probes on the substrate. Much as in a Southern hybridization, the target and oligonucleotide probes are of lengths typically greater than about 25 nucleotides. Under appropriate hybridization conditions, e.g., typically higher salt and lower temperature, the probes will hybridize irrespective of imperfect complementarity. In fact, with probes of greater than, e.g., about fifty nucleotides, the difference in stability of different sized probes will be relatively minor.

Typically the fingerprinting is merely for probing similarity or homology. Thus, the stringency of hybridization can usually be decreased to fairly low levels. See, e.g., Wetmur and Davidson (1968) "Kinetics of Renaturation of DNA," J. Mol. Biol., 31:349-370; and Kanehisa, M. (1984) Nuc. Acids Res., 12:203-213.

#### E. Detection; VLSIPS Scanning

Detection methods will be selected which are appropriate for the selected label. The scanning device need not necessarily be digitized or placed into a specific digital database, though such would most likely be done. For example, the analysis in fingerprinting could be photographic. Where a standardized fingerprint substrate matrix is used, the pattern of hybridizations may be spatially unique and may be compared photographically. In this manner, each sample may have a characteristic pattern of interactions and the likelihood of identical patterns will preferably be such low frequency that the fingerprint pattern indeed becomes a characteristic pattern virtually as unique as an individual's fingertip fingerprint. With a standardized substrate, every individual could be, in

WO 92/10588

PCT/US91/09226

41

theory, uniquely identifiable on the basis of the pattern of hybridizing to the substrate.

Of course, the VLSIPS scanning apparatus may also be useful to generate a digitized version of the fingerprint pattern. In this way, the identification pattern can be provided in a linear string of digits. This sequence could also be used for a standardized identification system providing significant useful medical transferability of specific data. In one embodiment, the probes used are selected to be of sufficiently high resolution to measure polynucleotides encoding antigens of the major histocompatibility complex, it might even be possible to provide transplantation matching data in a linear stream of data. The fingerprinting data may provide a condensed version, or summary, of the linear genetic data, or any other information data base.

#### F. Analysis

The analysis of the fingerprint will often be much simpler than a total sequence determination. However, there may be particular types of analysis which will be substantially simplified by a selected group of probes. For example, probes which exhibit particular populational heterogeneity may be selected. In this way, analysis may be simplified and practical utility enhanced merely by careful selection of the specific probes and a careful matrix layout of those probes.

#### G. Substrate Reuse

As with the sequencing application, the fingerprinting usages may also take advantage of the reusability of the substrate. In this way, the interactions can be disrupted, the substrate treated, and the renewed substrate is equivalent to an unused substrate.

#### H. Other Polynucleotide Aspects

Besides using the fingerprinting method for analyzing the structure of a particular polynucleotide, the fingerprinting method may be used to characterize various samples. For example, a cell or population of cells may be

WO 92/10588

PCT/US91/09226

42

tested for their expression of particular mRNA sequences, or for patterns of expressed mRNA species. This may be applicable to a cell or tissue type, to the expressed messenger RNA population expressed by a cell to the genetic content of a  
5 cell.

RNA can be isolated from a cell or a cell population, such as a purified cell fraction or a biopsy sample. The RNA may be labeled, for example by attaching a fluorescent molecule to isolated RNA or by using radiolabeled RNA (e.g., end-labeled  
10 with T4 polynucleotide kinase). A VLSIPS substrate containing positionally discrete oligonucleotide sequences may then be exposed to the pool of labeled RNA species under conditions permitting specific hybridization. The pattern of positions at which labeled RNA has formed specific hybrids may be compared  
15 to a reference pattern to identify, and in some embodiments quantify, the expressed RNA species, or to identify the hybridization pattern itself as being characteristic of a particular cell type.

For example but not for limitation, a VLSIPS  
20 oligonucleotide substrate may be hybridized to a labeled RNA sample obtained from a first cell type (e.g., human lymphocytes) to establish a reference hybridization pattern for the first cell type. Similarly, an identical VLSIPS oligonucleotide substrate may be hybridized to a labeled RNA  
25 sample obtained from a second cell type (e.g., human monocytes) to establish a reference hybridization pattern for the second cell type. Labeled RNA may then be prepared from a cell or a cell population and hybridized to an identical VLSIPS oligonucleotide substrate, and the resultant hybridization  
30 pattern can be compared to the reference hybridization patterns established for the first and second cell types. By such comparisons, the RNA expression pattern of a cell or cell population can be identified as being similar to or distinct from one or more reference hybridization patterns.

35 Where a positionally discrete oligonucleotide on the VLSIPS substrate is in molar excess over the amount of the cognate (complementary) labeled RNA species in the hybridization reaction, the amount of specific hybridization to

WO 92/10588

PCT/US91/09226

43

that VLSIPS locus (as measured by labeling intensity at that locus) can provide a quantitative measurement of the cognate RNA species present in the labeled RNA sample. Thus, hybridization of labeled RNA to a VLSIPS oligonucleotide substrate can provide information identifying the individual RNA species that are expressed in a particular cell or cell population, as well as the relative abundance of one or more individual RNA species. This information can serve to fingerprint specific cell types or particular stages in cell differentiation.

For example but not for limitation, RNA samples prepared from tissue biopsies, specifically tumor biopsies, can be labeled and hybridized to a VLSIPS oligonucleotide substrate, and the resultant hybridization pattern can provide information regarding cell type, degree of differentiation, and metastatic potential (malignancy). Some of the positionally distinct oligonucleotides may hybridize specifically with RNA species transcribed from endogenous proto-oncogens (e.g., c-myc, c-ras<sup>H</sup>, c-sis, etc.) which are, in certain instances, transcribed at elevated levels in neoplastic tissues.

In addition to diagnostic applications, labeled RNA samples from various neoplastic cell types may be hybridized to VLSIPS oligonucleotide substrates and the resultant hybridization pattern(s) compared to reference patterns obtained with RNA from related, non-neoplastic cell types. Identification of distinctions between the hybridization patterns obtained with RNA from neoplastic cells as compared to patterns obtained from RNA from non-neoplastic cells may be of diagnostic value and may identify RNA species that encode proteins that are potential targets for novel therapeutic modalities. In fact, the high resolution of the test will allow more complete characterization of parameters which define particular diseases. Thus, the power of diagnostic tests may be limited by the extent of statistical correlation with a particular condition rather than with the number of RNA species which are tested. The present invention provides the means to generate this large universe of possible reagents and the ability to actually accumulate that correlative data.

WO 92/10588

PCT/US91/09226

44

For fingerprinting of RNA expression patterns, the VLSIPS substrate polynucleotides will be at least 12 nucleotides in length, preferably at least 15 nucleotides in length, more preferably at least 25 nucleotides in length. The sequences of the positionally distinct polynucleotides on the VLSIPS substrate may be selected from published sources of sequence data, including but not limited to computerized database such as GenBank, and may or may not include random or pseudorandom sequences for detecting RNA species which have not yet been identified in the art. Fingerprint analysis of RNA expression patterns will typically employ high-stringency washes so as to provide hybridization patterns that reflect predominantly specific hybridization. However, some non-specific hybridization and/or cross-hybridization to slightly mismatched sequences may be tolerated, and in some embodiments may be desirable.

The ability to generate a high density means for screening the presence or absence of specific interactions allows for the possibility of screening for, if not saturating, all of a very large number of possible interactions. This is very powerful in providing the means for testing the combinations of molecular properties which can define a class of samples. For example, a species of organism may be characterized by its DNA sequences, e.g., a genetic fingerprint. By using a fingerprinting method, it may be determined that all members of that species are sufficiently similar in specific sequences that they can be easily identified as being within a particular group. Thus, newly defined classes may be resolved by their similarity in fingerprint patterns. Alternatively, a non-member of that group will fail to share those many identifying characteristics. However, since the technology allows testing of a very large number of specific interactions, it also provides the ability to more finely distinguish between closely related different cells or samples. This will have important applications in diagnosing viral, bacterial, and other pathological on nonpathological infections.

WO 92/10588

PCT/US91/09226

45

In particular, cell classification may be defined by any of a number of different properties. For example, a cell class may be defined by its DNA sequences contained therein. This allows species identification for parasitic or other

5 infections. For example, the human cell is presumably genetically distinguishable from a monkey cell, but different human cells will share many genetic markers. At higher resolution, each individual human genome will exhibit unique sequences that can define it as a single individual.

10 Likewise, a developmental stage of a cell type may be definable by its pattern of expression of messenger RNA. For example, in particular stages of cells, high levels of ribosomal RNA are found whereas relatively low levels of other types of messenger RNAs may be found. The high resolution

15 distinguishability provided by this fingerprinting method allows the distinction between cells which have relatively minor differences in its expressed mRNA population. Where a pattern is shown to be characteristic of a stage, a stage may be defined by that particular pattern of messenger RNA

20 expression.

In another embodiment, a substrate as provided herein may be used for genetic screening. This would allow for simultaneous screening of thousands of genetic markers. As the density of the matrix is increased, many more molecules can be

25 simultaneously tested. Genetic screening then becomes a simpler method as the present invention provides the ability to screen for thousands, tens of thousands, and hundreds of thousands, even millions of different possible genetic features. However, the number of high correlation genetic

30 markers for conditions numbers only in the hundreds. Again, the possibility for screening a large number of sequences provides the opportunity for generating the data which can provide correlation between sequences and specific conditions or susceptibility. The present invention provides the means to

35 generate extremely valuable correlations useful for the genetic detection of the causative mutation leading to medical conditions. In still another embodiment, the present invention would be applicable to distinguishing two individuals having

WO 92/10588

PCT/US91/09226

46

identical genetic compositions. The antibody population within an individual is dependent both on genetic and historical factors. Each individual experiences a unique exposure to various infectious agents, and the combined antibody expression  
5 is partly determined thereby. Thus, individuals may also be fingerprinted by their lymphocyte DNA or RNA hybridization pattern(s). Similar sorts of immunological and environmental histories may be useful for fingerprinting, perhaps in combination with other screening properties.

10 With the definition of new classes of cells, a cell sorter will be used to purify them. Moreover, new markers for defining that class of cells will be identified. For example, where the class is defined by its RNA content, cells may be screened by antisense probes which detect the presence or  
15 absence of specific sequences therein. Alternatively, cell lysates may provide information useful in correlating intracellular properties with extracellular markers which indicate functional differences. Using standard cell sorter technology with a fluorescence or labeled antisense probe which  
20 recognizes the internal presence of the specific sequences of interest, the cell sorter will be able to isolate a relatively homogeneous population of cells possessing the particular marker. Using successive probes the sorting process should be able to select for cells having a combination of a large number  
25 of different markers.

With the fingerprinted method as in identification means arises from mosaism problems in an organism. A mosaic organism is one whose genetic content in different cells is significantly different. Various clonal populations should  
30 have similar genetic fingerprints, though different clonal populations may have different genetic contents. See, for example, Suzuki et al. An Introduction to Genetic Analysis (4th Ed.), Freeman and Co., New York, which is hereby incorporated herein by reference. However, this problem should be a  
35 relatively rare problem and could be more carefully evaluated with greater experience using the fingerprinting methods.

The invention will also find use in detecting changes, both genetic and in protein expression (i.e., by RNA

WO 92/10588

PCT/US91/09226

47

expression fingerprinting), in a rapidly "evolving" protozoan infection, or similarly changing organism.

## V. MAPPING

### 5 A. General

The use of the present invention for mapping parallels its use for fingerprinting and sequencing. Mapping provides the ability to locate particular segments along the length of the polynucleotide. The mapping provides the ability  
10 to locate, in a relative sense, the order of various subsequences. This may be achieved using at least two different approaches.

The first approach is to take the large sequence and fragment it at specific points. The fragments are then ordered  
15 and attached to a solid substrate. For example, the clones resulting from a chromosome walking process may be individually attached to the substrate by methods, e.g., caged biotin techniques, indicated earlier. Segments of unknown map position will be exposed to the substrate and will hybridize to  
20 the segment which contains that particular sequence. This procedure allows the rapid determination of a number of different labeled segments, each mapping requiring only a single hybridization step once the substrate is generated. The substrate may be regenerated by removal of the interaction, and  
25 the next mapping segment applied.

In an alternative method, a plurality of subsequences can be attached to a substrate. Various short probes may be applied to determine which segments may contain particular overlaps. The theoretical basis and a description of this  
30 mapping procedure is contained in, e.g., Evans et al. 1989 "Physical Mapping of Complex Genomes by Cosmid Multiplex Analysis," Proc. Natl. Acad. Sci. USA 86:5030-5034, and other references cited above in the Section labeled "Overall Description." Using this approach, the details of the mapping  
35 embodiment are very similar to those used in the fingerprinting embodiment.

WO 92/10588

PCT/US91/09226

48

### B. Preparation of Substrate Matrix

The substrate may be generated in either of the methods generally applicable in the sequencing and fingerprinting embodiments. The substrate may be made either  
5 synthetically, or by attaching otherwise purified probes or sequences to the matrix. The probes or sequences may be derived either from synthetic or biological means. As indicated above, the solid phase substrate synthetic methods may be utilized to generate a matrix with positionally defined  
10 sequences. In the mapping embodiment, the importance of saturation of all possible subsequences of a preselected length is far less important than in the sequencing embodiment, but the length of the probes used may be desired to be much longer. The processes for making a substrate which has longer  
15 oligonucleotide probes should not be significantly different from those described for the sequencing embodiments, but the optimization parameters may be modified to comply with the mapping needs.

### 20 C. Labeling

The labeling methods will be similar to those applicable in sequencing and fingerprinting embodiments. Again, the target sequences may be desired to be fragmented.

### 25 D. Hybridization/Specific Interaction

The specificity of interaction between the targets and probe would typically be closer to those used for fingerprinting embodiments, where homology is more important than absolute distinguishability of high fidelity complementary  
30 hybridization. Usually, the hybridization conditions will be such that merely homologous segments will interact and provide a positive signal. Much like the fingerprinting embodiment, it may be useful to measure the extent of homology by successive incubations at higher stringency conditions. Or, a plurality  
35 of different probes, each having various levels of homology may be used. In either way, the spectrum of homologies can be measured.

WO 92/10588

PCT/US91/09226

49

#### E. Detection

The detection methods used in the mapping procedure will be virtually identical to those used in the fingerprinting  
5 embodiment. The detection methods will be selected in combination with the labeling methods.

#### F. Analysis

The analysis of the data in a mapping embodiment will  
10 typically be somewhat different from that in fingerprinting. The fingerprinting embodiment will test for the presence or absence of specific or homologous segments. However, in the mapping embodiment, the existence of an interaction is coupled with some indication of the location of the interaction. The  
15 interaction is mapped in some manner to the physical polymer sequence. Some means for determining the relative positions of different probes is performed. This may be achieved by synthesis of the substrate in pattern, or may result from analysis of sequences after they have been attached to the  
20 substrate.

For example, the probes may be randomly positioned at various locations on the substrate. However, the relative positions of the various reagents in the original polymer may be determined by using short fragments, e.g., individually, as  
25 target molecules which determine the proximity of different probes. By an automated system of testing each different short fragment of the original polymer, coupled with proper analysis, it will be possible to determine which probes are adjacent one another on the original target sequence and correlate that with  
30 positions on the matrix. In this way, the matrix is useful for determining the relative locations of various new segments in the original target molecule. This sort of analysis is described in Evans, and the related references described above.

In another form of mapping, as described above in the  
35 fingerprinting section, the developmental map of a cell or biological system may be measured using fingerprinting type technology. Thus, the mapping may be along a temporal dimension rather than along a polymer dimension. The mapping

WO 92/10588

PCT/US91/09226

50

or fingerprinting embodiments may also be used in determining the genetic rearrangements which may be genetically important, as in lymphocyte and B-cell development. In another example, various rearrangements or chromosomal dislocations may be tested by either the fingerprinting or mapping methods. These techniques are similar in many respects and the fingerprinting and mapping embodiments may overlap in many respects.

G. Substrate Reuse

The substrate should be reusable in the manner described in the fingerprinting section. The substrate is renewed by removal of the specific interactions and is washed and prepared for successive cycles of exposure to new target sequences.

VI. ADDITIONAL SCREENING AND APPLICATIONS

A. Specific Interactions

As originally indicated in the parent filing of VLSIPS, the production of a high density plurality of spatially segregated polymers provides the ability to generate a very large universe or repertoire of individually and distinct sequence possibilities. As indicated above, particular oligonucleotides may be synthesized in automated fashion at specific locations on a matrix. In fact, these oligonucleotides may be used to direct other molecules to specific locations by linking specific oligonucleotides to other reagents which are in batch exposed to the matrix and hybridized in a complementary fashion to only those locations where the complementary oligonucleotide has been synthesized on the matrix. This allows for spatially attaching a plurality of different reagents onto the matrix instead of individually attaching each separate reagent at each specific location. Although the caged biotin method allows the automated attachment, the speed of the caged biotin attachment process is relatively slow and requires a separate reaction for each reagent being attached. By use of the oligonucleotide method, the specificity of position can be done in an automated and parallel fashion. As each reagent is produced, instead of

WO 92/10588

PCT/US91/09226

51

directly attaching each reagent at each desired position, the reagent may be attached to a specific desired complementary oligonucleotide which will ultimately be specifically directed toward locations on the matrix having a complementary

5 oligonucleotide attached thereat.

In addition, the technology allows screening for specificity of interaction with particular reagents. For example, the oligonucleotide sequence specificity of binding of a potential reagent may be tested by presenting to the reagent all of the possible subsequences available for binding. Although secondary or higher order sequence specific features might not be easily screenable using this technology, it does provide a convenient, simple, quick, and thorough screen of interactions between a reagent and its target recognition sequences. See, e.g., Pfeifer et al. (1989) Science 246:810-812.

For example, the interaction of a promoter protein with its target binding sequence may be tested for many different, or all, possible binding sequences. By testing the strength of interactions under various different conditions, the interaction of the promoter protein with each of the different potential binding sites may be analyzed. The spectrum of strength of interactions with each different potential binding site may provide significant insight into the types of features which are important in determining specificity.

An additional example of a sequence specific interaction between reagents is the testing of binding of a double stranded nucleic acid structure with a single stranded oligonucleotide. Often, a triple stranded structure is produced which has significant aspects of sequence specificity. Testing of such interactions with either sequences comprising only natural nucleotides, or perhaps the testing of nucleotide analogs may be very important in screening for particularly useful diagnostic or therapeutic reagents. See, e.g., Häner and Dervan (1990) Biochemistry 29:9761-6765, and references therein.

WO 92/10588

PCT/US91/09226

52

### B. Sequence Comparisons

Once a gene is sequenced, the present invention provides means to compare alleles or related sequences to locate and identify differences from the control sequence.

- 5 This would be extremely useful in further analysis of genetic variability at a specific gene locus.

### C. Categorizations

- As indicated above in the fingerprinting and mapping  
10 embodiments, the present invention is also useful to define specific stages in the temporal sequence of cells, e.g., development, and the resulting tissues within an organism. For example, the developmental stage of a cell, or population of cells, can be dependent upon the expression of particular  
15 messenger RNAs. The screening procedures provided allow for high resolution definition of new classes of cells. In addition, the temporal development of particular cells will be characterized by the presence or expression of various mRNAs. Means to simultaneously screen a plurality or very large number  
20 of different sequences as provided. The combination of different markers made available dramatically increases the ability to distinguish fairly closely related cell types. Other markers may be combined with markers and methods made available herein to define new classifications of biological  
25 samples, e.g., based upon new combinations of markers.

- The presence or absence of particular marker sequences will be used to define temporal developmental stages. Once the stages are defined, fairly simple methods can be applied to actually purify those particular cells. For  
30 example, antisense probes or recognition reagents may be used with a cell sorter to select those cells containing or expressing the critical markers. Alternatively, the expression of those sequences may result in specific antigens which may also be used in defining cell classes and sorting those cells  
35 away from others. In this way, for example, it should be possible to select a class of omnipotent immune system cells which are able to completely regenerate a human immune system. Based upon the cellular classes defined by the parameters made

WO 92/10588

PCT/US91/09226

53

available by this technology, purified classes of cells having identifiable differences in RNA expression and/or DNA structure are made available.

In an alternative embodiment, subclasses of T-cells are defined, in part, upon the combination of expressed cell surface RNA species. The present invention allows for the simultaneous screening of a large plurality of different RNA species together. Thus, higher resolution classification of different T-cell subclasses becomes possible and, with the definitions and functional differences which correlate with those other parameters, the ability to purify those cell types becomes available. This is applicable not only to T-cells, lymphocyte cells, or even to freely circulating cells. Many of the cells for which this would be most useful will be immobile cells found in particular tissues or organs. Tumor cells will be diagnosed or detected using these fingerprinting techniques. Coupled with a temporal change in structure, developmental classes may also be selected and defined using these technologies. The present invention also provides the ability not only to define new classes of cells based upon functional or structural differences, but it also provides the ability to select or purify populations of cells which share these particular properties. In particular, antisense DNA or RNA molecules may be introduced into a cell to detect RNA sequences therein. See, e.g., Weintraub (1990) Scientific American 262:40-46.

#### D. Statistical Correlations

In an additional embodiment, the present invention also allows for the high resolution correlation of medical conditions with various different markers. For example, the present technology, when applied to amniocentesis or other genetic screening methods, typically screen for tens of different markers at most. The present invention allows simultaneous screening for tens, hundreds, thousands, tens of thousands, hundreds of thousands, and even millions of different genetic sequences. Thus, applying the fingerprinting methods of the present invention to a sufficiently large

WO 92/10588

PCT/US91/09226

54

population allows detailed statistical analysis to be made, thereby correlating particular medical conditions with particular markers, typically genetic markers or pathognomonic RNA expression patterns. Tumor-specific RNA expression patterns and particular RNA species characterizing various neoplastic phenotypes will be identified using the present invention.

Various medical conditions may be correlated against an enormous data base of the sequences within an individual. Genetic propensities and correlations then become available and high resolution genetic predictability and correlation become much more easily performed. With the enormous data base, the reliability of the predictions also is better tested. Particular markers which are partially diagnostic of particular medical conditions or medical susceptibilities will be identified and provide direction in further studies and more careful analysis of the markers involved. Of course, as indicated above in the sequencing embodiment, the present invention will find much use in intense sequencing projects. For example, sequencing of the entire human genome in the human genome project will be greatly simplified and enabled by the present invention.

#### VI. FORMATION OF SUBSTRATE

The substrate is provided with a pattern of specific reagents which are positionally localized on the surface of the substrate. This matrix of positions is defined by the automated system which produces the substrate. The instrument will typically be one similar to that described in PCT publication no. WO90/15070, and U.S.S.N. 07/624,120. The instrumentation described therein is directly applicable to the applications used here. In particular, the apparatus comprises a substrate, typically a silicon containing substrate, on which positions on the surface may be defined by a coordinate system of positions. These positions can be individually addressed or detected by the VLSIPS apparatus.

Typically, the VLSIPS apparatus uses optical methods used in semiconductor fabrication applications. In this way, masks may be used to photo-activate positions for attachment or

WO 92/10588

PCT/US91/09226

55

synthesis of specific sequences on the substrate. These manipulations may be automated by the types of apparatus described in PCT publication no. WO90/15070 and U.S.S.N. 07/624,120.

5           Selectively removable protecting groups allow creation of well defined areas of substrate surface having differing reactivities. Preferably, the protecting groups are selectively removed from the surface by applying a specific activator, such as electromagnetic radiation of a specific  
10 wavelength and intensity. More preferably, the specific activator exposes selected areas of surface to remove the protecting groups in the exposed areas.

          Protecting groups of the present invention are used in conjunction with solid phase oligonucleotide syntheses using  
15 deoxyribonucleic and ribonucleic acids. In addition to protecting the substrate surface from unwanted reaction, the protecting groups block a reactive end of the monomer to prevent self-polymerization.

          Attachment of a protecting group to the 5'-hydroxyl  
20 group of a nucleoside during synthesis using for example, phosphate-triester coupling chemistry, prevents the 5'-hydroxyl of one nucleoside from reacting with the 3'-activated phosphate-triester of another.

          Regardless of the specific use, protecting groups are  
25 employed to protect a moiety on a molecule from reacting with another reagent. Protecting groups of the present invention have the following characteristics: they prevent selected reagents from modifying the group to which they are attached; they are stable (that is, they remain attached) to the  
30 synthesis reaction conditions; they are removable under conditions that do not adversely affect the remaining structure; and once removed, do not react appreciably with the surface or surface-bound oligonucleotide.

          In a preferred embodiment, the protecting groups will  
35 be photoactivatable. The properties and uses of photoreactive protecting compounds have been reviewed. See, McCray *et al.*, Ann. Rev. of Biophys. and Biophys. Chem. (1989) 18:239-270, which is incorporated herein by reference. Preferably, the

WO 92/10588

PCT/US91/09226

56

photosensitive protecting groups will be removable by radiation in the ultraviolet (UV) or visible portion of the electromagnetic spectrum. More preferably, the protecting groups will be removable by radiation in the near UV or visible portion of the spectrum. In some embodiments, however, activation may be performed by other methods such as localized heating, electron beam lithography, laser pumping, oxidation or reduction with microelectrodes, and the like. Sulfonyl compounds are suitable reactive groups for electron beam lithography. Oxidative or reductive removal is accomplished by exposure of the protecting group to an electric current source, preferably using microelectrodes directed to the predefined regions of the surface which are desired for activation. A more detailed description of these protective groups is provided in U.S.S.N. 07/624,120, which is hereby incorporated herein by reference.

The density of reagents attached to a silicon substrate may be varied by standard procedures. The surface area for attachment of reagents may be increased by modifying the silicon surface. For example, a matte surface may be machined or etched on the substrate to provide more sites for attachment of the particular reagents. Another way to increase the density of reagent binding sites is to increase the derivitization density of the silicon. Standard procedures for achieving this are described, below.

One method to control the derivatization density is to highly derivatize the substrate with photochemical groups at high density. The substrate is then photolyzed for various predetermined times, which photoactivate the groups at a measurable rate, and react then with a capping reagent. By this method, the density of linker groups may be modulated by using a desired time and intensity of photoactivation.

In many applications, the number of different sequences which may be provided may be limited by the density and the size of the substrate on which the matrix pattern is generated. In situations where the density is insufficiently high to allow the screening of the desired number of sequences, multiple substrates may be used to increase the number of

WO 92/10588

PCT/US91/09226

57

sequences tested. Thus, the number of sequences tested may be increased by using a plurality of different substrates. Because the VLSIPS apparatus is almost fully automated, increasing the number of substrates does not lead to a significant increase in the number of manipulations which must be performed by humans. This again leads to greater reproducibility and speed in the handling of these multiple substrates.

10           A.   Instrumentation

The concept of using VLSIPS generally allows a pattern or a matrix of reagents to be generated. The procedure for making the pattern is performed by any of a number of different methods. An apparatus and instrumentation useful for generating a high density VLSIPS substrate is described in detail in PCT publication no. WO90/15070 and U.S.S.N. 07/624,120.

20           B.   Binary Masking

The details of the binary masking are described in an accompanying application filed simultaneously with this, U.S.S.N. 07/624,120, whose specification is incorporated herein by reference.

For example, the binary masking technique allows for producing a plurality of sequences based on the selection of either of two possibilities at any particular location. By a series of binary masking steps, the binary decision may be the determination, on a particular synthetic cycle, whether or not to add any particular one of the possible subunits. By treating various regions of the matrix pattern in parallel, the binary masking strategy provides the ability to carry out spatially addressable parallel synthesis.

35           C.   Synthetic Methods

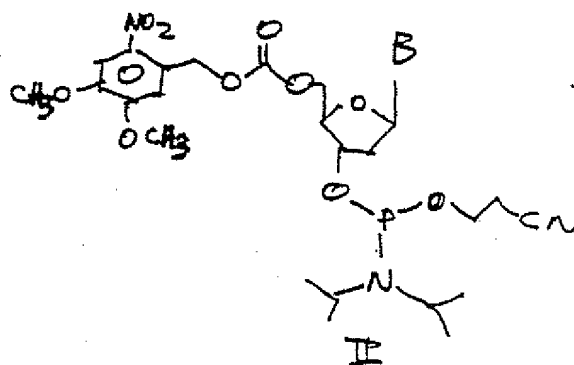
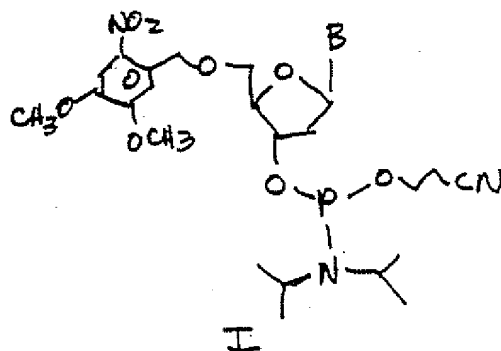
The synthetic methods in making a substrate are described in the parent application, U.S.S.N. 07/492,462. The construction of the matrix pattern on the substrate will typically be generated by the use of photo-sensitive reagents.

WO 92/10588

58

PCT/US91/09226

By use of photo-lithographic optical methods, particular segments of the substrate can be irradiated with light to activate or deactivate blocking agents, e.g., to protect or deprotect particular chemical groups. By an appropriate sequence of photo-exposure steps at appropriate times with appropriate masks and with appropriate reagents, the substrates can have known polymers synthesized at positionally defined regions on the substrate. Methods for synthesizing various substrates are described in PCT publication no. WO90/15070 and U.S.S.N. 07/624,120. By a sequential series of these photo-exposure and reaction manipulations, a defined matrix pattern of known sequences may be generated, and is typically referred to as a VLSIPS substrate. In the nucleic acid synthesis embodiment, nucleosides used in the synthesis of DNA by photolytic methods will typically be one of the two forms shown below:



B = Adenine, Cytosine, Guanine, or Thymine

WO 92/10588

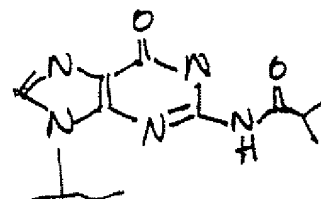
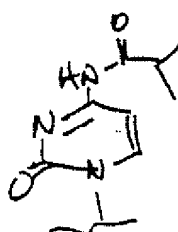
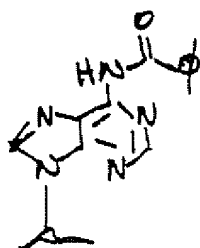
59

PCT/US91/09226

In I, the photolabile group at the 5' position is abbreviated NV (nitroveratryl) and in II, the group is abbreviated NVOC (nitroveratryl oxycarbonyl). Although not shown above, bases (adenine, cytosine, and guanine) contain exocyclic  $\text{NH}_2$  groups which must be protected during DNA synthesis. Thymine contains no exocyclic  $\text{NH}_2$  and therefore requires no protection. The standard protecting groups for these anaines are shown below:

10

15



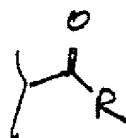
20 Adenine (A)

Cytosine (C)

Guanine (G)

Other amides of the general formula

25



R = alkyl  
aryl

30

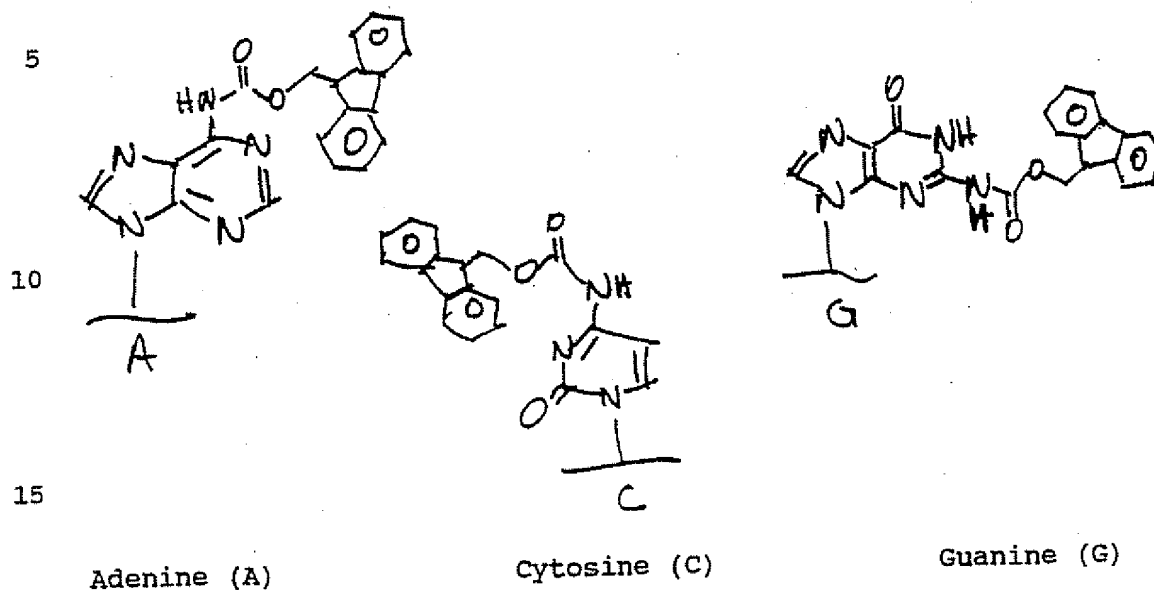
where R may be alkyl or aryl have been used.

WO 92/10588

60

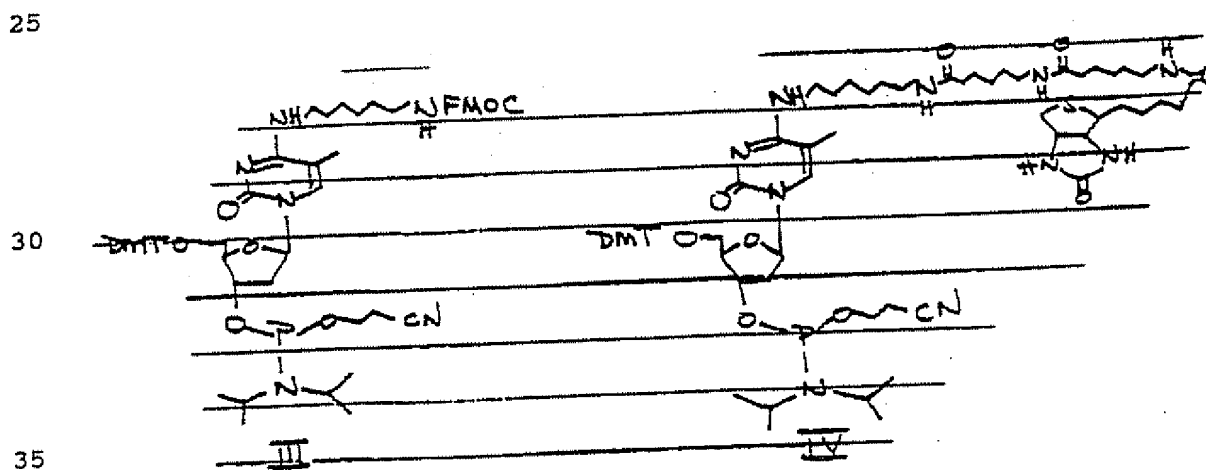
PCT/US91/09226

Another type of protecting group FMOC (9-fluorenyl methoxycarbonyl) is currently being used to protect the exocyclic amines of the three bases:



The advantage of the FMOC group is that it is removed under mild conditions (dilute organic bases) and can be used for all three bases. The amide protecting groups require more harsh conditions to be removed ( $\text{NH}_3/\text{MeOH}$  with heat).

Nucleosides used as 5'-OH probes, useful in verifying correct VLSIPS synthetic function, have been the following:



WO 92/10588

61

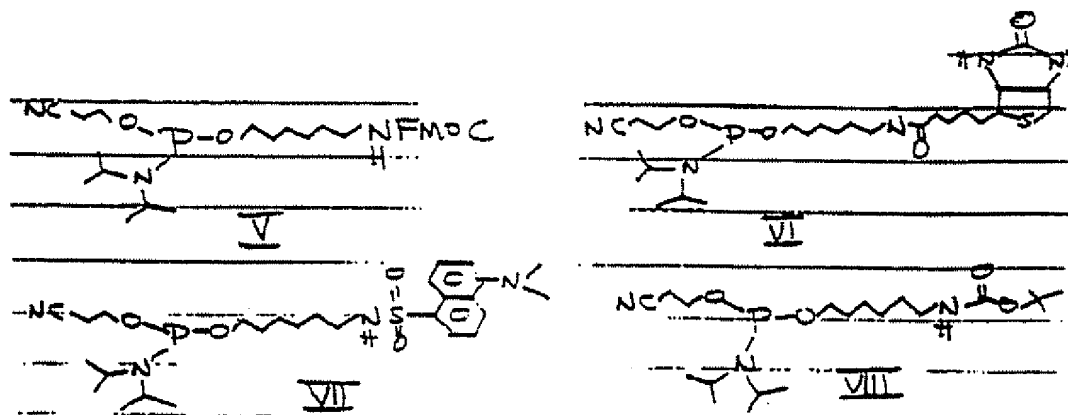
PCT/US91/09226

These compounds are used to detect where on a substrate photolysis has occurred by the attachment of either III or V to the newly generated 5'-OH. In the case of III, after the phosphate attachment is made, the substrate is

5 treated with a dilute base to remove the FMOC group. The resulting amine can be reacted with FITC and the substrate examined by fluorescence microscopy. This indicates the proper generation of a 5'-OH. In the case of compound IV, after the phosphate attachment is made, the substrate is treated with

10 FITC labeled streptavidin and the substrate again may be examined by fluorescence microscopy. Other probes, although not nucleoside based, have included the following:

15



20

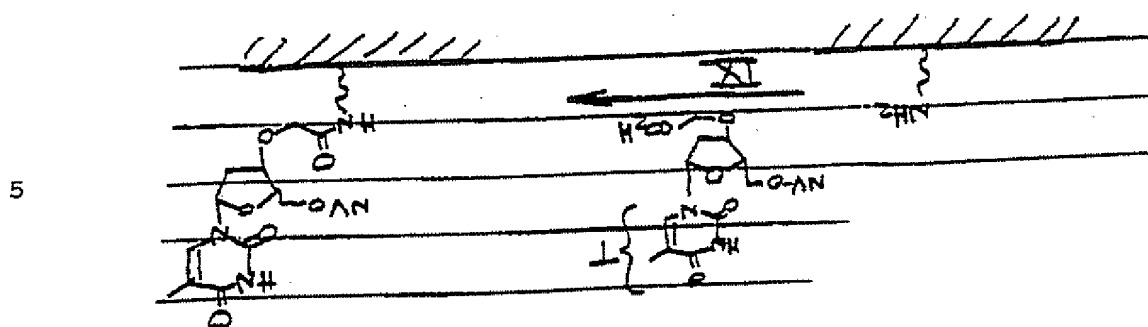
25

30 The method of attachment of the first nucleoside to the surface of the substrate depends on the functionality of the groups at the substrate surface. If the surface is amine functionalized, an amide bond is made (see example below).

WO 92/10588

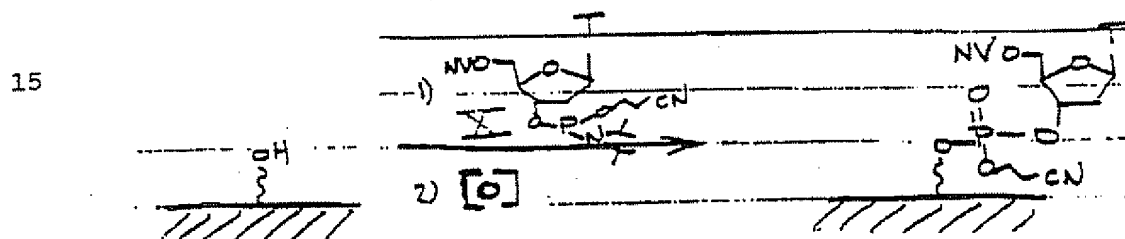
62

PCT/US91/09226



10

If the surface is hydroxy functionalized a phosphate bond is made (see example below)



20

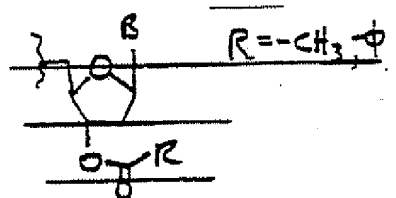
In both cases, the thymidine example is illustrated, but any one of the four phosphoramidite activated nucleosides can be used in the first step.

25 Photolysis of the photolabile group NV or NVOC on the 5' positions of the nucleosides is carried out at ~362 nm with an intensity of 14 mW/cm<sup>2</sup> for 10 minutes with the substrate side (side containing the photolabile group) immersed in dioxane. After the coupling of the next nucleoside is complete, the photolysis is repeated followed by another

30 coupling until the desired oligomer is obtained.

One of the most common 3'-O-protecting group is the ester, in particular the acetate

35



WO 92/10588

63

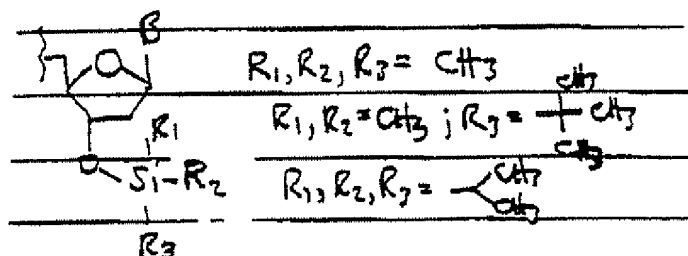
PCT/US91/09226

The groups can be removed by mild base treatment 0.1N NaOH/MeOH or  $K_2CO_3/H_2O/MeOH$ .

Another group used most often is the silyl ether.

5

10

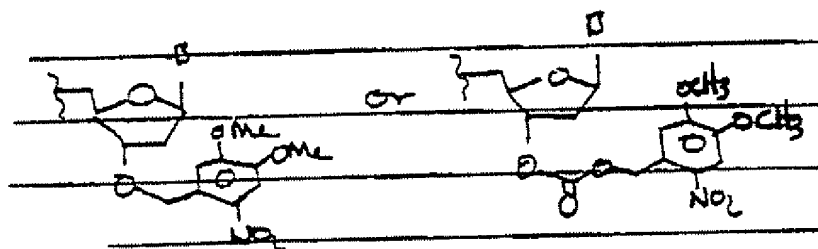


15

These groups can be removed by neutral conditions using 1 M tetra-n-butylammonium fluoride in THF or under acid conditions.

Related to photodeprotection, the nitroveratryl group could also be used to protect the 3'-position.

20

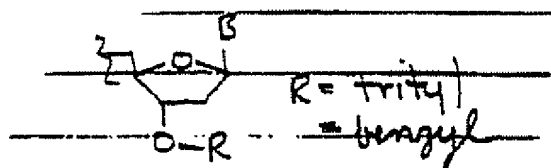


25

Here, light (photolysis) would be used to remove these protecting groups.

A variety of ethers can also be used in the protection of the 3'-O-position.

30



35

WO 92/10588

PCT/US91/09226

64

Removal of these groups usually involves acid or catalytic methods.

Note that corresponding linkages and photoblocked amino acids are described in detail in U.S.S.N. 07/624,120, which is hereby incorporated herein by reference.

Although the specificity of interactions at particular locations will usually be homogeneous due to a homogeneous polymer being synthesized at each defined location, for certain purposes, it may be useful to have mixed polymers with a commensurate mixed collection of interactions occurring at specific defined locations, or degeneracy reducing analogues, which have been discussed above and show broad specificity in binding. Then, a positive interaction signal may result from any of a number of sequences contained therein.

As an alternative method of generating a matrix pattern on a substrate, preformed polymers may be individually attached at particular sites on the substrate. This may be performed by individually attaching reagents one at a time to specific positions on the matrix, a process which may be automated. See, e.g., U.S.S.N. 07/435,316 (caged biotin parent), and U.S.S.N. 07/612,671 (caged biotin CIP). Another way of generating a positionally defined matrix pattern on a substrate is to have individually specific reagents which interact with each specific position on the substrate. For example, oligonucleotides may be synthesized at defined locations on the substrate. Then the substrate would have on its surface a plurality of regions having homogeneous oligonucleotides attached at each position.

In particular, at least four different substrate preparation procedures are available for treating a substrate surface. They are the standard VLSIPS method, polymeric substrates, Durapore™, and synthetic beads or fibers. The treatment labeled "standard VLSIPS" method is described in U.S.S.N. 07/624,120, and involves applying amino-propyltriethoxysilane to a glass surface.

The polymeric substrate approach involves either of two ways of generating a polymeric substrate. The first uses a high concentration of aminopropyltriethoxysilane (2-20%) in an

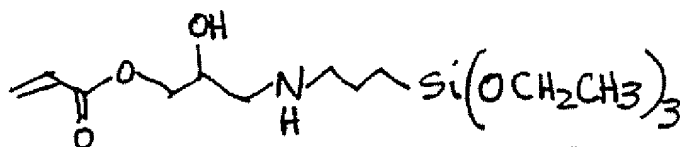
WO 92/10588

65

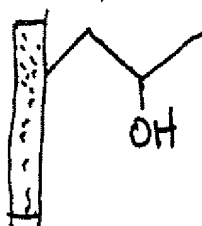
PCT/US91/09226

aqueous ethanol solution (95%). This allows the silane compound to polymerize both in solution and on the substrate surface, which provides a high density of amines on the surface of the glass. This density is contrasted with the standard VLSIPS method. This polymeric method allows for the deposition on the substrate surface of a monolayer due to the anhydrous method used with the aforementioned silane.

The second polymeric method involves either the coating or covalent binding of an appropriate acrylic acid polymer onto the substrate surface. In particular, e.g., in DNA synthesis, a monomer such as a hydroxypropylacrylate is used to generate a high density of hydroxyl groups on the substrate surface, allowing for the formation of phosphate bonds. An example of such a compound is shown:



The method using a Durapore™ membrane (Millipore) consists of a polyvinylidene difluoride coating with crosslinked polyhydroxylpropyl acrylate [PVDF-HPA]:



Here the building up of, e.g., a DNA oligomer, can be started immediately since phosphate bonds to the surface can be accomplished in the first step with no need for modification. A nucleotide dimer (5'-C-T-3') has been successfully made on this substrate in our labs.

The fourth method utilizes synthetic beads or fibers. This would use another substrate, such as a teflon copolymer

WO 92/10588

PCT/US91/09226

66

graft bead or fiber, which is covalently coated with an organic layer (hydrophilic) terminating in hydroxyl sites (commercially available from Molecular BroSystems, Inc.) This would offer the same advantage as the Durapore<sup>TM</sup> membrane, allowing for  
5 immediate phosphate linkages, but would give additional contour by the 3-dimensional growth of oligomers.

A matrix pattern of new reagents may be targeted to each specific oligonucleotide position by attaching a complementary oligonucleotide to which the substrate bound form  
10 is complementary. For instance, a number of regions may have homogeneous oligonucleotides synthesized at various locations. Oligonucleotide sequences complementary to each of these can be individually generated and linked to a particular specific reagents. Often these specific reagents will be antibodies.  
15 As each of these is specific for finding its complementary oligonucleotide, each of the specific reagents will bind through the oligonucleotide to the appropriate matrix position. A single step having a combination of different specific reagents being attached specifically to a particular  
20 oligonucleotide will thereby bind to its complement at the defined matrix position. The oligonucleotides will typically then be covalently attached, using, e.g., an acridine dye, for photocrosslinking. Psoralen is a commonly used acridine dye for photocrosslinking purposes, see, e.g., Song et al. (1979)  
25 Photochem. Photobiol. 29:1177-1197; Cimino et al. (1985) Ann. Rev. Biochem. 54:1151-1193; Parsons (1980) Photochem. Photobiol. 32:813-821; and Dattagupta et al. (1985) U.S. Pat. No. 4,542,102, and (1987) U.S. Pat. No. 4,713,326; each of which is hereby incorporated herein by reference. This method  
30 allows a single attachment manipulation to attach all of the specific reagents to the matrix at defined positions and results in the specific reagents being homogeneously located at defined positions.

35 D. Surface Immobilization

1. caged biotin

An alternative method of attaching reagents in a positionally defined matrix pattern is to use a caged biotin

WO 92/10588

PCT/US91/09226

67

system. See U.S.S.N. 07/612,671 (caged biotin CIP), which is hereby incorporated herein by reference, for additional details on the chemistry and application of caged biotin embodiments. In short, the caged biotin has a photosensitive blocking moiety which prevents the combination of avidin to biotin. At positions where the photo-lithographic process has removed the blocking group, high affinity biotin sites are generated. Thus, by a sequential series of photolithographic deblocking steps interspersed with exposure of those regions to appropriate biotin containing reagents, only those locations where the deblocking takes place will form an avidin-biotin interaction. Because the avidin-biotin binding is very tight, this will usually be virtually irreversible binding.

## 2. crosslinked interactions

The surface immobilization may also take place by photocrosslinking of defined oligonucleotides linked to specific reagents. After hybridization of the complementary oligonucleotides, the oligonucleotides may be crosslinked by a reagent by psoralen or another similar type of acridine dye. Other useful crosslinking reagents are described in Dattagupta et al. (1985) U.S. Pat. No. 4,542,102, and (1987) U.S. Pat. No. 4,713,326.

In another embodiment, colony or phage plaque transfer of biological polymers may be transferred directly onto a silicon substrate. For example, a colony plate may be transferred onto a substrate having a generic oligonucleotide sequence which hybridizes to another generic complementary sequence contained on all of the vectors into which inserts are cloned. This will specifically only bind those molecules which are actually contained in the vectors containing the desired complementary sequence. This immobilization allows for producing a matrix onto which a sequence specific reagent can bind, or for other purposes. In a further embodiment, a plurality of different vectors each having a specific oligonucleotide attached to the vector may be specifically attached to particular regions on a matrix having a complementary oligonucleotide attached thereto.

WO 92/10588

PCT/US91/09226

68

## VIII. HYBRIDIZATION/SPECIFIC INTERACTION

A. General

As discussed previously in the VLSIPS parent applications, the VLSIPS substrates may be used for screening  
5 for specific interactions with sequence specific targets or probes.

In addition, the availability of substrates having the entire repertoire of possible sequences of a defined length opens up the possibility of sequencing by hybridization. This  
10 sequence may be de novo determination of an unknown sequence, particularly of nucleic acid, verification of a sequence determined by another method, or an investigation of changes in a previously sequenced gene, locating and identifying specific changes. For example, often Maxam and Gilbert sequencing  
15 techniques are applied to sequences which have been determined by Sanger and Coulson. Each of those sequencing technologies have problems with resolving particular types of sequences. Sequencing by hybridization may serve as a third and independent method for verifying other sequencing techniques.  
20 See, e.g., (1988) Science 242:1245.

In addition, the ability to provide a large repertoire of particular sequences allows use of short subsequence and hybridization as a means to fingerprint a  
25 polynucleotide sample. For example, fingerprinting to a high degree of specificity of sequence matching may be used for identifying highly similar samples, e.g., those exhibiting high homology to the selected probes. This may provide a means for determining classifications of particular sequences. This should allow determination of whether particular genomes of  
30 bacteria, phage, or even higher cells might be related to one another.

In addition, fingerprinting may be used to identify an individual source of biological sample. See, e.g., Lander, E. (1989) Nature, 339:501-505, and references therein. For  
35 example, a DNA fingerprint may be used to determine whether a genetic sample arose from another individual. This would be particularly useful in various sorts of forensic tests to determine, e.g., paternity or sources of blood samples.

WO 92/10588

PCT/US91/09226

69

Significant detail on the particulars of genetic fingerprinting for identification purposes are described in, e.g., Morris et al. (1989) "Biostatistical evolution of evidence from continuous allele frequency distribution DNA probes in reference to disputed paternity of identity," J. Forensic Science 34:1311-1317; and Neufeld et al. (1990) Scientific American 262:46-53; each of which is hereby incorporated herein by reference.

In another embodiment, a fingerprinting-like procedure may be used for classifying cell types by analyzing a pattern of specific nucleic acids present in the cell, specifically RNA expression patterns. This may also be useful in defining the temporal stage of development of cells, e.g., stem cells or other cells which undergo temporal changes in development. For example, the stage of a cell, or group of cells, may be tested or defined by isolating a sample of mRNA from the population and testing to see what sequences are present in messenger populations. Direct samples, or amplified samples (e.g., by polymerase chain reaction), may be used. Where particular mRNA or other nucleic acid sequences may be characteristic of or shown to be characteristic of particular developmental stages, physiological states, or other conditions, this fingerprinting method may define them.

The present invention may also be used for mapping sequences within a larger segment. This may be performed by at least two methods, particularly in reference to nucleic acids. Often, enormous segments of DNA are subcloned into a large plurality of subsequences. Ordering these subsequences may be important in determining the overlaps of sequences upon nucleotide determinations. Mapping may be performed by immobilizing particularly large segments onto a matrix using the VLSIPS technology. Alternatively, sequences may be ordered by virtue of subsequences shared by overlapping segments. See, e.g., Craig et al. (1990) Nuc. Acids Res. 18:2653-2660; Michiels et al. (1987) CABIOS 3:203-210; and Olson et al. (1986) Proc. Natl. Acad. Sci. USA 83:7826-7830.

WO 92/10588

PCT/US91/09226

70

### B. Important Parameters

The extent of specific interaction between reagents immobilized to the VLSIPS substrate and another sequence specific reagent may be modified by the conditions of the interaction. Sequencing embodiments typically require high fidelity hybridization and the ability to discriminate perfect matching from imperfect matching. Fingerprinting and mapping embodiments may be performed using less stringent conditions, or in some embodiments very highly stringent conditions, depending upon the circumstances.

In a nucleic acid hybridization embodiment, the specificity and kinetics of hybridization have been described in detail by, e.g., Wetmur and Davidson (1968) J. Mol. Biol., 31:349-370, Britten and Kohne (1968) Science 161:529-530, and Kanehisa, (1984) Nuc. Acids Res. 12:203-213, each of which is hereby incorporated herein by reference. Parameters which are well known to affect specificity and kinetics of reaction include salt conditions, ionic composition of the solvent, hybridization temperature, length of oligonucleotide matching sequences, guanine and cytosine (GC) content, presence of hybridization accelerators, pH, specific bases found in the matching sequences, solvent conditions, and addition of organic solvents.

In particular, the salt conditions required for driving highly mismatched sequences to completion typically include a high salt concentration. The typical salt used is sodium chloride (NaCl), however, other ionic salts may be utilized, e.g., KCl. Depending on the desired stringency hybridization, the salt concentration will often be less than about 3 molar, more often less than 2.5 molar, usually less than about 2 molar, and more usually less than about 1.5 molar. For applications directed towards higher stringency matching, the salt concentrations would typically be lower. Ordinary high stringency conditions will utilize salt concentration of less than about 1 molar, more often less than about 750 millimolar, usually less than about 500 millimolar, and may be as low as about 250 or 150 millimolar.

WO 92/10588

PCT/US91/09226

71

The kinetics of hybridization and the stringency of hybridization both depend upon the temperature at which the hybridization is performed and the temperature at which the washing steps are performed. Temperatures at which steps for low stringency hybridization are desired would typically be lower temperatures, e.g., ordinarily at least about 15°C, more ordinarily at least about 20°C, usually at least about 25°C, and more usually at least about 30°C. For those applications requiring high stringency hybridization, or fidelity of hybridization and sequence matching, temperatures at which hybridization and washing steps are performed would typically be high. For example, temperatures in excess of about 35°C would often be used, more often in excess of about 40°C, usually at least about 45°C, and occasionally even temperatures as high as about 50°C or 60°C or more. Of course, the hybridization of oligonucleotides may be disrupted by even higher temperatures. Thus, for stripping of targets from substrates, as discussed below, temperatures as high as 80°C, or even higher may be used.

The base composition of the specific oligonucleotides involved in hybridization affects the temperature of melting, and the stability of hybridization as discussed in the above references. However, the bias of GC rich sequences to hybridize faster and retain stability at higher temperatures can be compensated for by the inclusion in the hybridization incubation or wash steps of various buffers. Sample buffers which accomplish this result include the triethyl- and trimethyl ammonium buffers. See, e.g., Wood et al. (1987) Proc. Natl. Acad. Sci. USA, 82:1585-1588, and Khrapko, K. et al. (1989) FEBS Letters 256:118-122.

The rate of hybridization can also be affected by the inclusion of particular hybridization accelerators. These hybridization accelerators include the volume exclusion agents characterized by dextran sulfate, or polyethylene glycol (PEG). Dextran sulfate is typically included at a concentration of between 1% and 40% by weight. The actual concentration selected depends upon the application, but typically a faster hybridization is desired in which the concentration is

WO 92/10588

PCT/US91/09226

72

optimized for the system in question. Dextran sulfate is often included at a concentration of between 0.5% and 2% by weight or dextran sulfate at a concentration between about 0.5% and 5%. Alternatively, proteins which accelerate hybridization may be added, e.g., the recA protein found in E. coli) or other homologous proteins.

Of course, the specific hybridization conditions will be selected to correspond to a discriminatory condition which provides a positive signal where desired but fails to show a positive signal at affinities where interaction is not desired. This may be determined by a number of titration steps or with a number of controls which will be run during the hybridization and/or washing steps to determine at what point the hybridization conditions have reached the stage of desired specificity.

#### IX. DETECTION METHODS

Methods for detection depend upon the label selected. The criteria for selecting an appropriate label are discussed below, however, a fluorescent label is preferred because of its extreme sensitivity and simplicity. Standard labeling procedures are used to determine the positions where interactions between a sequence and a reagent take place. For example, if a target sequence is labeled and exposed to a matrix of different probes, only those locations where probes do interact with the target will exhibit any signal. Alternatively, other methods may be used to scan the matrix to determine where interaction takes place. Of course, the spectrum of interactions may be determined in a temporal manner by repeated scans of interactions which occur at each of a multiplicity of conditions. However, instead of testing each individual interaction separately, a multiplicity of sequence interactions may be simultaneously determined on a matrix.

##### A. Labeling Techniques

The target polynucleotide may be labeled by any of a number of convenient detectable markers. A fluorescent label is preferred because it provides a very strong signal with low

WO 92/10588

PCT/US91/09226

73

background. It is also optically detectable at high resolution and sensitivity through a quick scanning procedure. Other potential labeling moieties include, radioisotopes, chemiluminescent compounds, labeled binding proteins, heavy  
5 metal atoms, spectroscopic markers, magnetic labels, and linked enzymes.

Another method for labeling does not require incorporation of a labeling moiety. The target may be exposed to the probes, and a double strand hybrid is formed at those  
10 positions only. Addition of a double strand specific reagent will detect where hybridization takes place. An intercalative dye such as ethidium bromide may be used as long as the probes themselves do not fold back on themselves to a significant extent forming hairpin loops. See, e.g., Sheldon et al. (1986)  
15 U.S. Pat. No. 4,582,789. However, the length of the hairpin loops in short oligonucleotide probes would typically be insufficient to form a stable duplex.

In another embodiment, different targets may be simultaneously sequenced where each target has a different  
20 label. For instance, one target could have a green fluorescent label and a second target could have a red fluorescent label. The scanning step will distinguish sites of binding of the red label from those binding the green fluorescent label. Each sequence can be analyzed independently from one another.

25 Suitable chromogens will include molecules and compounds which absorb light in a distinctive range of wavelengths so that a color may be observed, or emit light when irradiated with radiation of a particular wave length or wave length range, e.g., fluorescers.

30 A wide variety of suitable dyes are available, being primary chosen to provide an intense color with minimal absorption by their surroundings. Illustrative dye types include quinoline dyes, triarylmethane dyes, acridine dyes, alizarine dyes, phthaleins, insect dyes, azo dyes,  
35 anthraquinoid dyes, cyanine dyes, phenazathionium dyes, and phenazoxonium dyes.

A wide variety of fluorescers may be employed either by themselves or in conjunction with quencher molecules.

WO 92/10588

PCT/US91/09226

74

Fluorescers of interest fall into a variety of categories having certain primary functionalities. These primary functionalities include 1- and 2-aminonaphthalene, p,p'-diaminostilbenes, pyrenes, quaternary phenanthridine salts, 9-aminoacridines, p,p'-diaminobenzophenone imines, anthracenes, oxacarbocyanine, merocyanine, 3-aminoequilenin, perylene, bis-benzoxazole, bis-p-oxazolyl benzene, 1,2-benzophenazin, retinol, bis-3-aminopyridinium salts, hellebrigenin, tetracycline, sterophenol, benzimidzaolylphenylamine, 2-oxo-3-chromen, indole, xanthen, 7-hydroxycoumarin, phenoxazine, salicylate, strophanthidin, porphyrins, triarylmethanes and flavin. Individual fluorescent compounds which have functionalities for linking or which can be modified to incorporate such functionalities include, e.g., dansyl chloride; fluoresceins such as 3,6-dihydroxy-9-phenylxanthhydrol; rhodamineisothiocyanate; N-phenyl 1-amino-8-sulfonatonaphthalene; N-phenyl 2-amino-6-sulfonatonaphthalene; 4-acetamido-4-isothiocyanato-stilbene-2,2'-disulfonic acid; pyrene-3-sulfonic acid; 2-toluidinonaphthalene-6-sulfonate; N-phenyl, N-methyl 2-aminoaphthalene-6-sulfonate; ethidium bromide; stebrine; auromine-0,2-(9'-anthroyl)palmitate; dansyl phosphatidylethanolamine; N,N'-dioctadecyl oxacarbocyanine; N,N'-dihexyl oxacarbocyanine; merocyanine, 4-(3'pyrenyl)butyrate; d-3-aminodesoxy-equilenin; 12-(9'-anthroyl)stearate; 2-methylantracene; 9-vinylanthracene; 2,2'-(vinylene-p-phenylene)bisbenzoxazole; p-bis[2-(4-methyl-5-phenyl-oxazolyl)]benzene; 6-dimethylamino-1,2-benzophenazin; retinol; bis(3'-aminopyridinium) 1,10-decandiyl diiodide; sulfonaphthylhydrazone of hellibrienin; chlorotetracycline; N-(7-dimethylamino-4-methyl-2-oxo-3-chromenyl)maleimide; N-[p-(2-benzimidazolyl)-phenyl]maleimide; N-(4-fluoranthyl)maleimide; bis(homovanillic acid); resazarin; 4-chloro-7-nitro-2,1,3-benzooxadiazole; merocyanine 540; resorufin; rose bengal; and 2,4-diphenyl-3(2H)-furanone.

Desirably, fluorescers should absorb light above about 300 nm, preferably about 350 nm, and more preferably above about 400 nm, usually emitting at wavelengths greater than about 10 nm higher than the wavelength of the light

WO 92/10588

PCT/US91/09226

75

absorbed. It should be noted that the absorption and emission characteristics of the bound dye may differ from the unbound dye. Therefore, when referring to the various wavelength ranges and characteristics of the dyes, it is intended to indicate the dyes as employed and not the dye which is unconjugated and characterized in an arbitrary solvent.

Fluorescers are generally preferred because by irradiating a fluorescer with light, one can obtain a plurality of emissions. Thus, a single label can provide for a plurality of measurable events.

Detectable signal may also be provided by chemiluminescent and bioluminescent sources. Chemiluminescent sources include a compound which becomes electronically excited by a chemical reaction and may then emit light which serves as the detectible signal or donates energy to a fluorescent acceptor. A diverse number of families of compounds have been found to provide chemiluminescence under a variety of conditions. One family of compounds is 2,3-dihydro-1,4-phthalazinedione. The most popular compound is luminol, which is the 5-amino compound. Other members of the family include the 5-amino-6,7,8-trimethoxy- and the dimethylamino[ca]benz analog. These compounds can be made to luminesce with alkaline hydrogen peroxide or calcium hypochlorite and base. Another family of compounds is the 2,4,5-triphenylimidazoles, with lophine as the common name for the parent product. Chemiluminescent analogs include para-dimethylamino and -methoxy substituents. Chemiluminescence may also be obtained with oxalates, usually oxalyl active esters, e.g., p-nitrophenyl and a peroxide, e.g., hydrogen peroxide, under basic conditions. Alternatively, luciferins may be used in conjunction with luciferase or lucigenins to provide bioluminescence.

Spin labels are provided by reporter molecules with an unpaired electron spin which can be detected by electron spin resonance (ESR) spectroscopy. Exemplary spin labels include organic free radicals, transitional metal complexes, particularly vanadium, copper, iron, and manganese, and the like. Exemplary spin labels include nitroxide free radicals.

WO 92/10588

PCT/US91/09226

76

### B. Scanning System

With the automated detection apparatus, the correlation of specific positional labeling is converted to the presence on the target of sequences for which the reagents have specificity of interaction. Thus, the positional information is directly converted to a database indicating what sequence interactions have occurred. For example, in a nucleic acid hybridization application, the sequences which have interacted between the substrate matrix and the target molecule can be directly listed from the positional information. The detection system used is described in PCT publication no. WO90/15070; and U.S.S.N. 07/624,120. Although the detection described therein is a fluorescence detector, the detector may be replaced by a spectroscopic or other detector. The scanning system may make use of a moving detector relative to a fixed substrate, a fixed detector with a moving substrate, or a combination. Alternatively, mirrors or other apparatus can be used to transfer the signal directly to the detector. See, e.g., U.S.S.N. 07/624,120, which is hereby incorporated herein by reference.

The detection method will typically also incorporate some signal processing to determine whether the signal at a particular matrix position is a true positive or may be a spurious signal. For example, a signal from a region which has actual positive signal may tend to spread over and provide a positive signal in an adjacent region which actually should not have one. This may occur, e.g., where the scanning system is not properly discriminating with sufficiently high resolution in its pixel density to separate the two regions. Thus, the signal over the spatial region may be evaluated pixel by pixel to determine the locations and the actual extent of positive signal. A true positive signal should, in theory, show a uniform signal at each pixel location. Thus, processing by plotting number of pixels with actual signal intensity should have a clearly uniform signal intensity. Regions where the signal intensities show a fairly wide dispersion, may be

WO 92/10588

PCT/US91/09226

77

particularly suspect and the scanning system may be programmed to more carefully scan those positions.

In another embodiment, as the sequence of a target is determined at a particular location, the overlap for the sequence would necessarily have a known sequence. Thus, the system can compare the possibilities for the next adjacent position and look at these in comparison with each other. Typically, only one of the possible adjacent sequences should give a positive signal and the system might be programmed to compare each of these possibilities and select that one which gives a strong positive. In this way, the system can also simultaneously provide some means of measuring the reliability of the determination by indicating what the average signal to background ratio actually is.

More sophisticated signal processing techniques can be applied to the initial determination of whether a positive signal exists or not. See, e.g., U.S.S.N. 07/624,120.

From a listing of those sequences which interact, data analysis may be performed on a series of sequences. For example, in a nucleic acid sequence application, each of the sequences may be analyzed for their overlap regions and the original target sequence may be reconstructed from the collection of specific subsequences obtained therein. Other sorts of analyses for different applications may also be performed, and because the scanning system directly interfaces with a computer the information need not be transferred manually. This provides for the ability to handle large amounts of data with very little human intervention. This, of course, provides significant advantages over manual manipulations. Increased throughput and reproducibility is thereby provided by the automation of vast majority of steps in any of these applications.

#### DATA ANALYSIS

35

##### A. General

Data analysis will typically involve aligning the proper sequences with their overlaps to determine the target sequence. Although the target "sequence" may not specifically

WO 92/10588

PCT/US91/09226

78

correspond to any specific molecule, especially where the target sequence is broken and fragmented up in the sequencing process, the sequence corresponds to a contiguous sequence of the subfragments.

5           The data analysis can be performed by a computer using an appropriate program. See, e.g., Drmanac, R. et al. (1989) Genomics 4:114-128; and a commercially available analysis program available from the Genetic Engineering Center, P.O. Box 794, 11000 Belgrade, Yugoslavia. Although the

10 specific manipulations necessary to reassemble the target sequence from fragments may take many forms, one embodiment uses a sorting program to sort all of the subsequences using a defined hierarchy. The hierarchy need not necessarily

15 correspond to any physical hierarchy, but provides a means to determine, in order, which subfragments have actually been found in the target sequence. In this manner, overlaps can be checked and found directly rather than having to search

20 throughout the entire set after each selection process. For example, where the oligonucleotide probes are 10-mers, the first 9 positions can be sorted. A particular subsequence can be selected as in the examples, to determine where the process starts. As analogous to the theoretical example provided

25 above, the sorting procedure provides the ability to immediately find the position of the subsequence which contains the first 9 positions and can compare whether there exists more than 1 subsequence during the first 9 positions. In fact, the computer can easily generate all of the possible target sequences which contain given combination of subsequences. Typically there will be only one, but in various situations,

30 there will be more.

          An exemplary flow chart for a sequencing program is provided in Figure 4. In general terms, the program provides for automated scanning of the substrate to determine the positions of probe and target interaction. Simple processing

35 of the intensity of the signal may be incorporated to filter out clearly spurious signals. The positions with positive interaction are correlated with the sequence specificity of specific matrix positions, to generate the set of matching

WO 92/10588

PCT/US91/09226

79

subsequences. This information is further correlated with other target sequence information, e.g., restriction fragment analysis. The sequences are then aligned using overlap data, thereby leading to possible corresponding target sequences which will, optimally, correspond to a single target sequence.

#### B. Hardware

A variety of computer systems may be used to run a sequencing program. The program may be written to provide both the detecting and scanning steps together and will typically be dedicated to a particular scanning apparatus. However, the components and functional steps may be separated and the scanning system may provide an output, e.g., through tape or an electronic connection into a separate computer which separately runs the sequencing analysis program. The computer may be any of a number of machines provided by standard computer manufacturers, e.g., IBM compatible machines, Apple<sup>TM</sup> machines, VAX machines, and others, which may often use a UNIX<sup>TM</sup> operating system. Alternatively, custom computing architectures may be employed, these architectures may include neural network methods implemented in hardware and/or software. Of course, the hardware used to run the analysis program will typically determine what programming language would be used.

#### C. Software

Software would be readily developed by a person of ordinary skill in the programming art, following the flow chart provided, or based upon the input provided and the desired result.

Of course, an exemplary embodiment is a polynucleotide sequence system. However, the theoretical and mathematical manipulations necessary for data analysis of other linear molecules are conceptually similar.

#### XI. SUBSTRATE REUSE

Where a substrate is made with specific reagents that are relatively insensitive to the handling and processing steps involved in a single cycle of use, the substrate may often be

WO 92/10588

PCT/US91/09226

80

reused. The target molecules are usually stripped off of the solid phase specific recognition molecules. Of course, it is preferred that the manipulations and conditions be selected as to be mild and to not affect the substrate. For example, if a substrate is acid labile, a neutral pH would be preferred in all handling steps. Similar sensitivities would be carefully respected where recycling is desired.

#### A. Removal of Label

Typically for a recycling, the previously attached specific interaction would be disrupted and removed. This will typically involve exposing the substrate to conditions under which the interaction between probe and target is disrupted. Alternatively, it may be exposed to conditions where the target is destroyed. For example, where the probes are oligonucleotides and the target is a polynucleotide, a heating and low salt wash will often be sufficient to disrupt the interactions. Additional reagents may be added such as detergents, and organic or inorganic solvents which disrupt the interaction between the specific reagents and target.

#### B. Storage and Preservation

As indicated above, the matrix will typically be maintained under conditions where the matrix itself and the linkages and specific reagents are preserved. Various specific preservatives may be added which prevent degradation. For example, if the reagents are acid or base labile, a neutral pH buffer will typically be added. It is also desired to avoid destruction of the matrix by growth of organisms which may destroy organic reagents attached thereto. For this reason, a preservative such as cyanide or azide may be added. However, the chemical preservative should also be selected to preserve the chemical nature of the linkages and other components of the substrate. Typically, a detergent may also be included.

#### C. Processes to Avoid Degradation of Oligomers

In particular, a substrate comprising a large number of oligomers will be treated in a fashion which is known to

WO 92/10588

PCT/US91/09226

81

maintain the quality and integrity of oligonucleotides. These include storing the substrate in a carefully controlled environment under conditions of lower temperature, cation depletion (EDTA and EGTA), sterile conditions, and inert argon or nitrogen atmosphere.

## XII. INTEGRATED SEQUENCING STRATEGY

### A. Initial Mapping Strategy

As indicated above, although the VLSIPS may be applied to sequencing embodiments, it is often useful to integrate other concepts to simplify the sequencing. For example, nucleic acids may be easily sequenced by careful selection of the vectors and hosts used for amplifying and generating the specific target sequences. For example, it may be desired to use specific vectors which have been designed to interact most efficiently with the VLSIPS substrate. This is also important in fingerprinting and mapping strategies. For example, vectors may be carefully selected having particular complementary sequences which are designed to attach to a genetic or specific oligomer on the substrate. This is also applicable to situations where it is desired to target particular sequences to specific locations on the matrix.

In one embodiment, unnatural oligomers may be used to target natural probes to specific locations on the VLSIPS substrate. In addition, particular probes may be generated for the mapping embodiment which are designed to have specific combinations of characteristics. For example, the construction of a mapping substrate may depend upon use of another automated apparatus which takes clones isolated from a chromosome walk and attaches them individually or in bulk to the VLSIPS substrate.

In another embodiment, a variety of specific vectors having known and particular "targeting" sequences adjacent the cloning sites may be individually used to clone a selected probe, and the isolated probe will then be targetable to a site on the VLSIPS substrate with a sequence complementary to the "target" sequence.

WO 92/10588

PCT/US91/09226

82

### B. Selection of Smaller Clones

In the fingerprinting and mapping embodiments, the selection of probes may be very important. Significant mathematical analysis may be applied to determine which specific sequences should be used as those probes. Of course, for fingerprinting use, sequences that show significant heterogeneity across the human population would be preferred. Selection of the specific sequences which would most favorably be utilized will tend to be single copy sequences within the genome, and more specifically single copy sequences that have low cross-hybridization potential to other sequences in the genome (i.e., not members of a closely-related multigene family).

Various hybridization selection procedures may be applied to select sequences which tend not to be repeated within a genome, and thus would tend to be conserved across individuals. For example, hybridization selections may be made for non-repetitive and single copy sequences. See, e.g., Britten and Kohne (1968) "Repeated Sequences in DNA," Science 161:529-540. On the other hand, it may be desired under certain circumstances to use repeated sequences. For example, where a fingerprint may be used to identify or distinguish different species, or where repetitive sequences may be diagnostic of specific species, repetitive sequences may be desired for inclusion in the fingerprinting probes. In either case, the sequencing capability will greatly assist in the selection of appropriate sequences to be used as probes.

Also as indicated above, various means for constructing an appropriate substrate may involve either mechanical or automated procedures. The standard VLSIPS automated procedure involves synthesizing oligonucleotides or short polymers directly on the substrate. In various other embodiments, it is possible to attach separately synthesized reagents onto the matrix in an ordered array. Other circumstances may lend themselves to transfer a pattern from a petri plate onto a solid substrate. Also, there are methods for site specifically directing collections of reagents to

WO 92/10588

PCT/US91/09226

83

specific locations using unnatural nucleotides or equivalent sorts of targeting molecules.

While a brute force manual transfer process may be utilized sequentially attaching various samples to successive positions, instrumentation for automating such procedures may also be devised. The automated system for performing such would preferably be relatively easily designed and conceptually easily understood.

XIII. COMMERCIAL APPLICATIONS

A. Sequencing

As indicated above, sequencing may be performed either de novo or as a verification of another sequencing method. The present hybridization technology provides the ability to sequence nucleic acids and polynucleotides de novo, or as a means to verify either the Maxam and Gilbert chemical sequencing technique or Sanger and Coulson dideoxy- sequencing techniques. The hybridization method is useful to verify sequencing determined by any other sequencing technique and to closely compare two similar sequences, e.g., to identify and locate sequence differences.

Of course, sequencing of can be very important in many different sorts of environments. For example, it will be useful in determining the genetic sequence of particular markers in various individuals. In addition, polymers may be used as markers or for information containing molecules to encode information. For example, a short polynucleotide sequence may be included in large bulk production samples indicating the manufacturer, date, and location of manufacture of a product. For example, various drugs may be encoded with this information with a small number of molecules in a batch. For example, a pill may have somewhere from 10 to 100 to 1,000 or more very short and small molecules encoding this information. When necessary, this information may be decoded from a sample of the material using a polymerase chain reaction (PCR) or other amplification method. This encoding system may be used to provide the origin of large bulky samples without significantly affecting the properties of those samples. For

WO 92/10588

PCT/US91/09226

84

example, chemical samples may also be encoded by this method thereby providing means for identifying the source and manufacturing details of lots. The origin of bulk hydrocarbon samples may be encoded. Production lots of organic compounds  
5 such as benzene or plastics may be encoded with a short molecule polymer. Food stuffs may also be encoded using similar marking molecules. Even toxic waste samples can be encoded determining the source or origin. In this way, proper disposal can be traced or more easily enforced.

10 Similar sorts of encoding may be provided by fingerprinting-type analysis. Whether the resolution is absolute or less so, the concept of coding information on molecules such as nucleic acids, which can be amplified and later decoded, may be a very useful and important application.

15 This technology also provides the ability to include markers for origins of biological materials. For example, a patented animal line may be transformed with a particular unnatural sequence which can be traced back to its origin. With a selection of multiple markers, the likelihood could be  
20 negligible that a combination of markers would have independently arisen from a source other than the patented or specifically protected source. This technique may provide a means for tracing the actual origin of particular biological materials. Bacteria, plants, and animals will be subject to  
25 marking by such encoding sequences.

#### B. Fingerprinting

As indicated above, fingerprinting technology may also be used for data encryption. Moreover, fingerprinting  
30 allows for significant identification of particular individuals. Where the fingerprinting technology is standardized, and used for identification of large numbers of people, related equipment and peripheral processing will be developed to accompany the underlying technology. For example,  
35 specific equipment may be developed for automatically taking a biological sample and generating or amplifying the information molecules within the sample to be used in fingerprinting analysis. Moreover, the fingerprinting substrate may be mass

WO 92/10588

PCT/US91/09226

85

produced using particular types of automatic equipment. Synthetic equipment may produce the entire matrix simultaneously by stepwise synthetic methods as provided by the VLSIPS technology. The attachment of specific probes onto a substrate may also be automated, e.g., making use of the caged biotin technology. See, e.g., U.S.S.N. 07/612,671 (caged biotin CIP).

In addition, peripheral processing may be important and may be dedicated to this specific application. Thus, automated equipment for producing the substrates may be designed, or particular systems which take in a biological sample and output either a computer readout or an encoded instrument, e.g., a card or document which indicates the information and can provide that information to others. An identification having a short magnetic strip with a few million bits may be used to provide individual identification and important medical information useful in a medical emergency.

In fact, data banks may be set up to correlate all of this information of fingerprinting with medical information. This may allow for the determination of correlations between various medical problems and specific DNA sequences. By collating large populations of medical records with genetic information, genetic propensities and genetic susceptibilities to particular medical conditions may be developed. Moreover, with standardization of substrates, the micro encoding data may be also standardized to reproduce the information from a centralized data bank or on an encoding device carried on an individual person. On the other hand, if the fingerprinting procedure is sufficiently quick and routine, every hospital may routinely perform a fingerprinting operation and from that determine many important medical parameters for an individual.

In particular industries, the VLSIPS sequencing, fingerprinting, or mapping technology will be particularly appropriate. As mentioned above, agricultural livestock suppliers may be able to encode and determine whether their particular strains are being used by others. By incorporating particular markers into their genetic stocks, the markers will indicate origin of genetic material. This is applicable to

WO 92/10588

PCT/US91/09226

86

seed producers, livestock producers, and other suppliers of medical or agricultural biological materials.

This may also be useful in identifying individual animals or plants. For example, these markers may be useful in determining whether certain fish return to their original breeding grounds, whether sea turtles always return to their original birthplaces, or to determine the migration patterns and viability of populations of particular endangered species. It would also provide means for tracking the sources of particular animal products. For example, it might be useful for determining the origins of controlled animal substances such as elephant ivory or particular bird populations whose importation or exportation is controlled.

As indicated above, polymers may be used to encode important information on source and batch and supplier. This is described in greater detail, e.g., "Applications of PCR to industrial problems," (1990) in Chemical and Engineering News 68:145, which is hereby incorporated herein by reference. In fact, the synthetic method can be applied to the storage of enormous amounts of information. Small substrates may encode enormous amounts of information, and its recovery will make use of the inherent replication capacity. For example, on regions of  $10\text{ }\mu\text{m} \times 10\text{ }\mu\text{m}$ ,  $1\text{ cm}^2$  has  $10^6$  regions. An theory, the entire human genome could be attached in 1000 nucleotide segments on a  $3\text{ cm}^2$  surface. Genomes of endangered species may be stored on these substrates.

Fingerprinting may also be used for genetic tracing or for identifying individuals for forensic science purposes. See, e.g., Morris, J. et al. (1989) "Biostatistical Evaluation of Evidence From Continuous Allele Frequency Distribution DNA Probes in Reference to Disputed Paternity and Identity," J. Forensic Science 34:1311-1317, and references provided therein; each of which is hereby incorporated herein by reference.

In addition, the high resolution fingerprinting allows the distinguishability to high resolution of particular samples. As indicated above, new cell classifications may be defined based on combinations of a large number of properties. Similar applications will be found in distinguishing different

WO 92/10588

PCT/US91/09226

87

species of animals or plants. In fact, microbial identification may become dependent on characterization of the genetic content. Tumors or other cells exhibiting abnormal physiology will be detectable by use of the present invention.

- 5 Also, knowing the genetic fingerprint of a microorganism may provide very useful information on how to treat an infection by such organism.

- Modifications of the fingerprint embodiments may be used to diagnose the condition of the organism. For example, a  
10 blood sample is presently used for diagnosing any of a number of different physiological conditions. A multi-dimensional fingerprinting method made available by the present invention could become a routine means for diagnosing an enormous number of physiological features simultaneously. This may  
15 revolutionize the practice of medicine in providing information on an enormous number of parameters together at one time. In another way, the genetic predisposition may also revolutionize the practice of medicine providing a physician with the ability to predict the likelihood of particular medical conditions  
20 arising at any particular moment. It also provides the ability to apply preventative medicine.

- Also available are kits with the reagents useful for performing sequencing, fingerprinting, and mapping procedures. The kits will have various compartments with the desired  
25 necessary reagents, e.g., substrate, labeling reagents for target samples, buffers, and other useful accompanying products.

### C. Mapping

- 30 The present invention also provides the means for mapping sequences within enormous stretches of sequence. For example, nucleotide sequences may be mapped within enormous chromosome size sequence maps. For example, it would be possible to map a chromosomal location within the chromosome  
35 which contains hundreds of millions of nucleotide base pairs. In addition, the mapping and fingerprinting embodiments allow for testing of chromosomal translocations, one of the standard problems for which amniocentesis is performed.

WO 92/10588

PCT/US91/09226

88

The present invention will be better understood by reference to the following illustrative examples. The following examples are offered by way of illustration and not by way of limitation.

- 5           Relevant applications whose techniques are incorporated herein by reference are PCT publication no. WO90/15070, published December 13, 1990; PCT publication no. WO91/07087, published May 30, 1991; U.S.S.N. 07/624,120, filed December 6, 1990; and U.S.S.N. 07/626,730, filed December 6, 10   1990.
- Also, additional relevant techniques are described, e.g., in Sambrook, J., et al. (1989) Molecular Cloning: a Laboratory Manual, 2d Ed., vols 1-3, Cold Spring Harbor Press, New York; Greenstein and Winitz (1961) Chemistry of the Amino 15 Acids, Wiley and Sons, New York; Bodzansky, M. (1988) Peptide Chemistry: a Practical Textbook, Springer-Verlag, New York; Harlow and Lane (1988) Antibodies: A Laboratory Manual, Cold Spring Harbor Press, New York; Glover, D. (ed.) (1987) DNA Cloning: A Practical Approach, vols 1-3, IRL Press, Oxford; 20 Bishop and Rawlings (1987) Nucleic Acid and Protein Sequence Analysis: A Practical Approach, IRL Press, Oxford; Hames and Higgins (1985) Nucleic Acid Hybridisation: A Practical Approach, IRL Press, Oxford; Wu et al. (1989) Recombinant DNA Methodology, Academic Press, San Diego; Goding (1986) 25 Monoclonal Antibodies: Principles and Practice, (2d ed.), Academic Press, San Diego; Finegold and Barron (1986) Bailey and Scott's Diagnostic Microbiology, (7th ed.), Mosby Co., St. Louis; Collins et al. (1989) Microbiological Methods, (6th ed.), Butterworth, London; Chaplin and Kennedy (1986) 30 Carbohydrate Analysis: A Practical Approach, IRL Press, Oxford; Van Dyke (ed.) (1985) Bioluminescence and Chemiluminescence: Instruments and Applications, vol 1, CRC Press, Boca Rotan; and Ausubel et al. (ed.) (1990) Current Protocols in Molecular Biology, Greene Publishing and Wiley-Interscience, New York; 35 each of which is hereby incorporated herein by reference.

WO 92/10588

89

PCT/US91/09226

EXAMPLES

The following examples are provided to illustrate the efficacy of the inventions herein. All operations were conducted at about ambient temperatures and pressures unless indicated to the contrary.

## POLYNUCLEOTIDE SEQUENCING

## 1. HPLC of the photolysis of 5'-O-nitroveratryl-thymidine.

In order to determine the time for photolysis of 5'-O-nitroveratryl thymidine to thymidine a 100  $\mu$ M solution of NV-Thym-OH (5'-O-nitroveratryl thymidine) in dioxane was made and ~200  $\mu$ l aliquots were irradiated (in a quartz cuvette 1 cm x 2 mm) at 362.3 nm for 20 sec, 40 sec, 60 sec, 2 min, 5 min, 10 min, 15 min, and 20 min. The resulting irradiated mixtures were then analyzed by HPLC using a Varian MicroPak SP column (C<sub>18</sub> analytical) at a flow rate of 1 ml/min and a solvent system of 40% CH<sub>3</sub>CN and 60% water. Thymidine has a retention time of 1.2 min and NVO-Thym-OH has a retention time of 2.1 min. It was seen that after 10 min of exposure the deprotection was complete.

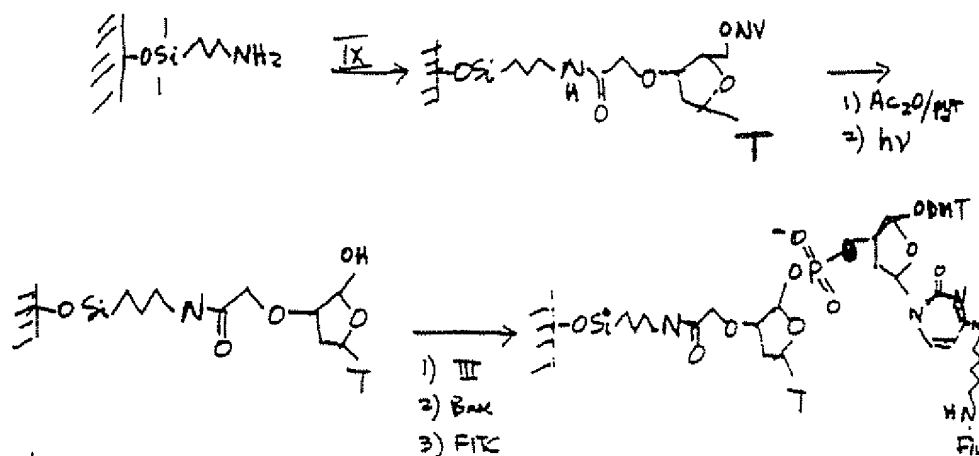
## 2. Preparation and Detection of Thymidine-Cytidine dimer (FITC)

The reaction is illustrated:

25

30

35



WO 92/10588

90

PCT/US91/09226 .

To an aminopropylated glass slide (standard VLSIPS) was added a mixture of the following:

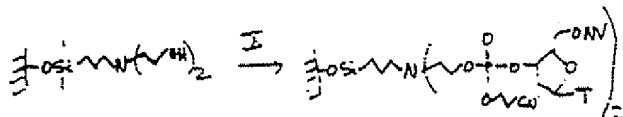
- 12.2 mg of NVO-Thym-CO<sub>2</sub>H (IX)
- 3.4 mg of HOBT (N-hydroxybenztriazal)
- 8.8  $\mu$ l DIEA (Diisopropylethylamine)
- 11.1 mg BOP reagent
- 2.5 ml DMF

After 2 h coupling time (standard VLSIPS) the plate was washed, acetylated with acetic anhydride/pyridine, washed, dried, and photolyzed in dioxane at 362 nm at 14 mW/cm<sup>2</sup> for 10 min using a 500  $\mu$ m checkerboard mask. The slide was then taken and treated with a mixture of the following:

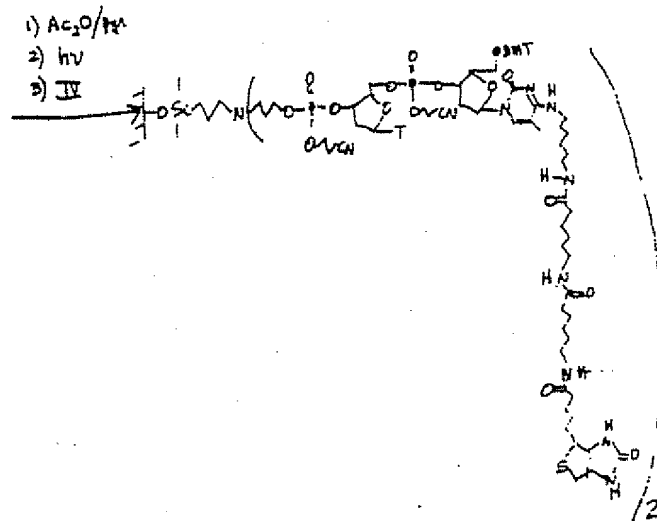
- 107 mg of FMOC-amine modified C (III)
- 21 mg of tetrazole
- 1 ml anhydrous CH<sub>3</sub>CN

After being treated for approximately 8 min, the slide was washed off with CH<sub>3</sub>CN, dried, and oxidized with I<sub>2</sub>/H<sub>2</sub>O/THF/lutidine for 1 min. The slide was again washed, dried, and treated for 30 min with a 20% solution of DBU in DMF. After thorough rinsing of the slide, it was next exposed to a FITC solution (1mM fluorescein isothiocyanate [FITC] in DMF) for 50 min, then washed, dried, and examined by fluorescence microscopy. This reaction is illustrated:

25



30



35

WO 92/10588

PCT/US91/09226

91

### 3. Preparation and Detection of Thymidine-cytidine dimer (Biotin)

An aminopropyl glass slide, was soaked in a solution of ethylene oxide (20% in DMF) to generate a hydroxylated surface. The slide was added a mixture of the following:

- 32 mg of NVO-T-OCED (X)
- 11 mg of tetrazole
- 0.5 ml of anhydrous  $\text{CH}_3\text{CN}$

After 8 min the plate was then rinsed with acetonitrile, then oxidized with  $\text{I}_2/\text{H}_2\text{O}/\text{THF}/\text{lutidine}$  for 1 min, washed and dried. The slide was then exposed to a 1:3 mixture of acetic anhydride:pyridine for 1 h, then washed and dried. The substrate was a then photolyzed in dioxane at 362 nm at 14  $\text{mW}/\text{cm}^2$  for 10 min using a 500  $\mu\text{m}$  checkerboard mask, dried, and then treated with a mixture of the following:

- 65 mg of biotin modified C (IV)
- 11 mg of tetrazole
- 0.5 ml anhydrous  $\text{CH}_3\text{CN}$

After 8 min the slide was washed with  $\text{CH}_3\text{CN}$  then oxidized with  $\text{I}_2/\text{H}_2\text{O}/\text{THF}/\text{lutidine}$  for 1 min, washed, and then dried. The slide was then soaked for 30 min in a PBS/0.05% Tween 20 buffer and the solution then shaken off. The slide was next treated with FITC-labeled streptavidin at 10  $\mu\text{g}/\text{ml}$  in the same buffer system for 30 min. After this time the streptavidin-buffer system was rinsed off with fresh PBS/0.05% Tween 20 buffer and then the slide was finally agitated in distilled water for about 1/2 h. After drying, the slide was examined by fluorescence microscopy (see Fig. 2 and Fig. 3).

### 4. substrate preparation

Before attachment of reactive groups it is preferred to clean the substrate which is, in a preferred embodiment, a glass substrate such as a microscope slide or cover slip. A roughened surface will be useable but a plastic or other solid substrate is also appropriate. According to one embodiment the slide is soaked in an alkaline bath consisting of, e.g., 1 liter of 95% ethanol with 120 ml of water and 120 grams of sodium hydroxide for 12 hours. The slides are washed with a

WO 92/10588

PCT/US91/09226

92

buffer and under running water, allowed to air dry, and rinsed with a solution of 95% ethanol.

The slides are then aminated with, e.g., aminopropyltriethoxysilane for the purpose of attaching amino groups to the glass surface on linker molecules, although other omega functionalized silanes could also be used for this purpose. In one embodiment 0.1% aminopropyltriethoxysilane is utilized, although solutions with concentrations from  $10^{-7}\%$  to 10% may be used, with about  $10^{-3}\%$  to 2% preferred. A 0.1% mixture is prepared by adding to 100 ml of a 95% ethanol/5% water mixture, 100 microliters ( $\mu$ l) of aminopropyltriethoxysilane. The mixture is agitated at about ambient temperature on a rotary shaker for an appropriate amount of time, e.g., about 5 minutes. 500  $\mu$ l of this mixture is then applied to the surface of one side of each cleaned slide. After 4 minutes or more, the slides are decanted of this solution and thoroughly rinsed three times or more by dipping in 100% ethanol.

After the slides dry, they are heated in a 110-120°C vacuum oven for about 20 minutes, and then allowed to cure at room temperature for about 12 hours in an argon environment. The slides are then dipped into DMF (dimethylformamide) solution, followed by a thorough washing with methylene chloride.

25

#### 5. linker attachment, blocking of free sites

The aminated surface of the slide is then exposed to about 500  $\mu$ l of, for example, a 30 millimolar (mM) solution of NVOC-nucleotide- NHS (N-hydroxysuccinimide) in DMF for attachment of a NVOC-nucleotide to each of the amino groups. See, e.g., SIGMA Chemical Company for various nucleotide derivatives. The surface is washed with, for example, DMF, methylene chloride, and ethanol.

Any unreacted aminopropyl silane on the surface, i.e., those amino groups which have not had the NVOC-nucleotide attached, are now capped with acetyl groups (to prevent further reaction) by exposure to a 1:3 mixture of acetic anhydride in pyridine for 1 hour. Other materials which may perform this

35

WO 92/10588

PCT/US91/09226

93

residual capping function include trifluoroacetic anhydride, formicacetic anhydride, or other reactive acylating agents. Finally, the slides are washed again with DMF, methylene chloride, and ethanol.

5

6. synthesis of eight trimers of C and T

Fig. 4 illustrates a possible synthesis of the eight trimers of the two-monomer set: cytosine and thymine (represented by C and T, respectively). A glass slide bearing silane groups terminating in 6-nitroveratryloxycarboxamide (NVOC-NH) residues is prepared as a substrate. Active esters (pentafluorophenyl, OBt, etc.) of cytosine and thymine protected at the 5' hydroxyl group with NVOC are prepared as reagents. While not pertinent to this example, if side chain protecting groups are required for the monomer set, these must not be photoreactive at the wavelength of light used to protect the primary chain.

For a monomer set of size  $n$ ,  $n \times \ell$  cycles are required to synthesize all possible sequences of length  $\ell$ . A cycle consists of:

1. Irradiation through an appropriate mask to expose the 5'-OH groups at the sites where the next residue is to be added, with appropriate washes to remove the by-products of the deprotection.
2. Addition of a single activated and protected (with the same photochemically-removable group) monomer, which will react only at the sites addressed in step 1, with appropriate washes to remove the excess reagent from the surface.

The above cycle is repeated for each member of the monomer set until each location on the surface has been extended by one residue in one embodiment. In other embodiments, several residues are sequentially added at one location before moving on to the next location. Cycle times will generally be limited by the coupling reaction rate, now as short as about 10 min in automated oligonucleotide synthesizers. This step is optionally followed by addition of

WO 92/10588

PCT/US91/09226 .

94

a protecting group to stabilize the array for later testing. For some types of polymers (e.g., peptides), a final deprotection of the entire surface (removal of photoprotective side chain groups) may be required.

5           More particularly, as shown in Fig. 4A, the glass 20 is provided with regions 22, 24, 26, 28, 30, 32, 34, and 36. Regions 30, 32, 34, and 36 are masked, indicated by the hatched regions, as shown in Fig. 4B and the glass is irradiated by the bright regions 22, 24, 26, and 28, and exposed to a reagent  
10   containing a photosensitive blocked C (e.g., cytosine derivative), with the resulting structure shown in Fig. 4C. The substrate is carefully washed and the reactants removed. Thereafter, regions 22, 24, 26, and 28 are masked, as indicated  
15   by the hatched region, the glass is irradiated (as shown in Fig. 4D), as indicated by the bright regions, at 30, 32, 34, and 36, and exposed to a photosensitive blocked reagent containing T (e.g., thymine derivative), with the resulting  
20   structure shown in Fig. 4E. The process proceeds, consecutively masking and exposing the sections as shown until the structure shown in Fig. 4M is obtained. The glass is irradiated and the terminal groups are, optionally, capped by acetylation. As shown, all possible trimers of  
cytosine/thymine are obtained.

25           In this example, no side chain protective group removal is necessary, as might be common in modified nucleotides. If it is desired, side chain deprotection may be accomplished by treatment with ethanedithiol and trifluoroacetic acid.

30           In general, the number of steps needed to obtain a particular polymer chain is defined by:

$$n \times \ell \quad (1)$$

where:

n = the number of monomers in the basis set of monomers, and

35           ℓ = the number of monomer units in a polymer chain.

Conversely, the synthesized number of sequences of length ℓ will be:

$$n^{\ell} \quad (2)$$

WO 92/10588

PCT/US91/09226

95

Of course, greater diversity is obtained by using masking strategies which will also include the synthesis of polymers having a length of less than  $\ell$ . If, in the extreme case, all polymers having a length less than or equal to  $\ell$  are synthesized, the number of polymers synthesized will be:

$$n^{\ell} + n^{\ell-1} + \dots + n^1. \quad (3)$$

The maximum number of lithographic steps needed will generally be  $n$  for each "layer" of monomers, i.e., the total number of masks (and, therefore, the number of lithographic steps) needed will be  $n \times \ell$ . The size of the transparent mask regions will vary in accordance with the area of the substrate available for synthesis and the number of sequences to be formed. In general, the size of the synthesis areas will be:

$$\text{size of synthesis areas} = (A)/(S)$$

where:

A is the total area available for synthesis; and  
S is the number of sequences desired in the area.

It will be appreciated by those of skill in the art that the above method could readily be used to simultaneously produce thousands or millions of oligomers on a substrate using the photolithographic techniques disclosed herein. Consequently, the method results in the ability to practically test large numbers of, for example, di, tri, tetra, penta, hexa, hepta, octa, nona, deca, even dodecanucleotides, or larger polynucleotides.

The above example has illustrated the method by way of a manual example. It will of course be appreciated that automated or semi-automated methods could be used. The substrate would be mounted in a flow cell for automated addition and removal of reagents, to minimize the volume of reagents needed, and to more carefully control reaction conditions. Successive masks will be applicable manually or automatically. See, e.g., PCT publication no. WO90/15070 and U.S.S.N. 07/624,120.

WO 92/10588

PCT/US91/09226

96

## 7. labeling of target

The target oligonucleotide can be labeled using standard procedures referred to above. As discussed, for certain situations, a reagent which recognizes interaction, e.g., ethidium bromide, may be provided in the detection step. Alternatively, fluorescence labeling techniques may be applied, see, e.g., Smith, et al. (1986) Nature, 321: 674-679; and Prober, et al. (1987) Science, 238:336-341. The techniques described therein will be followed with minimal modifications as appropriate for the label selected.

## 8. dimers of A, C, G, and T

The described technique may be applied, with photosensitive blocked nucleotides corresponding to adenine, cytosine, guanine, and thymine, to make combinations of polynucleotides consisting of each of the four different nucleotides. All 16 possible dimers would be made using a minor modification of the described method.

## 9. 10-mers of A, C, G, and T

The described technique for making dimers of A, C, G, and T may be further extended to make longer oligonucleotides. The automated system described, e.g., in PCT publication no. WO90/15070, and U.S.S.N. 07/624,120, can be adapted to make all possible 10-mers composed of the 4 nucleotides A, C, G, and T. The photosensitive, blocked nucleotide analogues have been described above, and would be readily adaptable to longer oligonucleotides.

## 10. specific recognition hybridization to 10-mers

The described hybridization conditions are directly applicable to the sequence specific recognition reagents attached to the substrate, produced as described immediately above. The 10-mers have an inherent property of hybridizing to a complementary sequence. For optimum discrimination between full matching and some mismatch, the conditions of hybridization should be carefully selected, as described above. Careful control of the conditions, and titration of parameters

WO 92/10588

PCT/US91/09226

97

should be performed to determine the optimum collective conditions.

#### 11. hybridization

5 Hybridization conditions are described in detail, e.g., in Hames and Higgins (1985) Nucleic Acid Hybridisation: A Practical Approach; and the considerations for selecting particular conditions are described, e.g., in Wetmur and Davidson, (1988) J. Mol. Biol. 31:349-370, and Wood et al. 10 (1985) Proc. Natl. Acad. Sci. USA 82:1585-1588. As described above, conditions are desired which can distinguish matching along the entire length of the probe from where there is one or more mismatched bases. The length of incubation and conditions will be similar, in many respects, to the hybridization 15 conditions used in Southern blot transfers. Typically, the GC bias may be minimized by the introduction of appropriate concentrations of the alkylammonium buffers, as described above.

Titration of the temperature and other parameters is 20 desired to determine the optimum conditions for specificity and distinguishability of absolutely matched hybridization from mismatched hybridization.

A fluorescently labeled target or set of targets are generated, as described in Prober, et al. (1987) Science 25 238:336-341, or Smith, et al. (1986) Nature 321:674-679. Preferably, the target or targets are of the same length as, or slightly longer, than the oligonucleotide probes attached to the substrate and they will have known sequences. Thus, only a few of the probes hybridize perfectly with the target, and 30 which particular ones did would be known.

The substrate and probes are incubated under appropriate conditions for a sufficient period of time to allow hybridization to completion. The time is measured to determine when the probe-target hybridizations have reached completion. 35 A salt buffer which minimizes GC bias is preferred, incorporating, e.g., buffer, such as tetramethyl ammonium or tetraethyl ammonium ion at between about 2.4 and 3.0 M. See Wood, et al. (1985) Proc. Nat'l Acad. Sci. USA 82:1585-1588.

WO 92/10588

PCT/US91/09226

98

This time is typically at least about 30 min, and may be as long as about 1-5 days. Typically very long matches will hybridize more quickly, very short matches will hybridize less quickly, depending upon relative target and probe concentrations. The hybridization will be performed under conditions where the reagents are stable for that time duration.

Upon maximal hybridization, the conditions for washing are titrated. Three parameters initially titrated are time, temperature, and cation concentration of the wash step. The matrix is scanned at various times to determine the conditions at which the distinguishability between true perfect hybrid and mismatched hybrid is optimized. These conditions will be preferred in the sequencing embodiments.

15

#### 12. positional detection of specific interaction

As indicated above, the detection of specific interactions may be performed by detecting the positions where the labeled target sequences are attached. Where the label is a fluorescent label, the apparatus described, e.g., PCT publication no. WO90/15070; and U.S.S.N. 07/624,120, may be advantageously applied. In particular, the synthetic processes described above will result in a matrix pattern of specific sequences attached to the substrate, and a known pattern of interactions can be converted to corresponding sequences.

In an alternative embodiment, a separate reagent which differentially interacts with the probe and interacted probe/targets can indicate where interaction occurs or does not occur. A single-strand specific reagent will indicate where no interaction has taken place, while a double-strand specific reagent will indicate where interaction has taken place. An intercalating dye, e.g., ethidium bromide, may be used to indicate the positions of specific interaction.

35

#### 13. analysis

Conversion of the positional data into sequence specificity will provide the set of subsequences whose analysis by overlap segments, may be performed, as described above.

WO 92/10588

PCT/US91/09226

99

Analysis is provided by the methodology described above, or using, e.g., software available from the Genetic Engineering Center, P.O. Box 794, 11000 Belgrade, Yugoslavia (Yugoslav group). See, also, Macevicz, PCT publication no. WO 90/04652, which is hereby incorporated herein by reference.

The description of the preparation of short peptides on a substrate incorporates by reference sections in U.S.S.N. 07/492,462 (VLSIPS CIP), and described below.

#### 10 POLYNUCLEOTIDE FINGERPRINTING

The above section on generation of reagents for sequencing provides specific reagents useful for fingerprinting applications. Fingerprinting embodiments may be applied towards polynucleotide fingerprinting, cell and tissue classification, cell and tissue temporal development stage classification, diagnostic tests, forensic uses for individual identification, classification of organisms, and genetic screening of individuals. Mapping applications are also described below.

20 Polynucleotide fingerprinting may use reagents similar to those described above for probing a sequence for the presence of specific subsequences found therein. Typically, the subsequences used for fingerprinting will be longer than the sequences used in oligonucleotide sequencing. In particular, specific long segments may be used to determine the similarity of different samples of nucleic acids. They may also be used to fingerprint whether specific combinations of information are provided therein. Particular probe sequences are selected and attached in a positional manner to a substrate. The means for attachment may be either using a caged biotin method described, e.g., in U.S.S.N. 07/612,671 (caged biotin CIP), or by another method using targeting molecules. In one embodiment, an unnatural nucleotide or similar complementary binding molecule may be attached to the fingerprinting probe and the probe thereby directed towards complementary sequences on a VLSIPS substrate. Typically, unnatural nucleotides would be preferred, e.g., unnatural

WO 92/10588

PCT/US91/09226 .

100

optical isomers, which would not interfere with natural nucleotide interactions.

Having produced a substrate with particular fingerprint probes attached thereto at positionally defined regions, the substrate may be used in a manner quite similar to the sequencing embodiment to provide information as to whether the fingerprint probes are detecting the corresponding sequence in a target sequence. This will often provide information similar to a Southern blot hybridization.

10

#### Temporal Development

Developmental RNA expression patterns

The present fingerprinting invention also allows cell classification by identification of developmental RNA expression patterns. For example, a lymphocyte stem cell expresses a particular combination of RNA species. As the lymphocyte develops through a program developmental scheme, at various stages it expresses particular RNA species which are diagnostic of particular stages in development. Again, the fingerprinting methodology allows for the definition of specific structural features which are diagnostic of developmental or functional features which will allow classification of cells into temporal developmental classes. Cells, products of those cells, or lysates of those cells will be assayed to determine the developmental stage of the source cells. In this manner, once a developmental stage is defined, specific synchronized populations of cells will be selected out of another population. These synchronized populations may be very important in determining the biological mechanisms of development.

30

The present invention also allows for fingerprinting of the mRNA population of a cell. In this fashion, the mRNA population, which should be a good determinant of developmental stage, will be correlated with other structural features of the cell. In this manner, cells at specific developmental stages will be characterized by the intracellular environment, as well as the extracellular environment.

35

WO 92/10588

PCT/US91/09226

101

### Diagnostic Tests

The present invention also provides the ability to perform diagnostic tests. Diagnostic tests typically are based upon a fingerprint type assay, which tests for the presence of specific diagnostic polynucleotides. Thus, the present invention provides means for viral strain identification, bacterial strain identification, and other diagnostic tests using positionally defined specific oligonucleotide reagents.

### Viral Identification

The present invention provides reagents and methodology for identifying viral strains. The viral genome may be probed for specific sequences which are characteristic of particular viral strains. Specific hybridization patterns on an VLSIPS oligonucleotide substrate can identify the presence of particular viral genomes.

### Bacterial Identification

Similar techniques will be applicable to identifying a bacterial source. This may be useful in diagnosing bacterial infections, or in classifying sources of particular bacterial species. For example, the bacterial assay may be useful in determining the natural range of survivability of particular strains of bacteria across regions of the country or in different ecological niches.

### Other Microbiological Identifications

The present invention provides means for diagnosis of other microbiological and other species, e.g., protozoal species and parasitic species in a biological sample, but also provides the means for assaying a combination of different infections. For example, a biological specimen may be assayed for the presence of any or all of these microbiological species. In human diagnostic uses, typical samples will be blood, sputum, stool, urine, or other samples.

WO 92/10588

PCI/US91/09226 .

102

### Individual Identification

The present invention provides the ability to fingerprint and identify a genetic individual. This individual may be a bacterial or lower microorganism, as described above  
5 in diagnostic tests, or of a plant or animal. An individual may be identified genetically, as described.

Genetic fingerprinting has been utilized in comparing different related species in Southern hybridization blots. Genetic fingerprinting has also been used in forensic studies,  
10 see, e.g., Morris et al. (1989) J. Forensic Science 34: 1311-1317, and references cited therein. As described above, an individual may be identified genetically by a sufficiently large number of probes. The likelihood that another individual  
15 would have an identical pattern over a sufficiently large number of probes may be statistically negligible. However, it is often quite important that a large number of probes be used where the statistical probability of matching is desired to be particularly low. In fact, the probes will optimally be selected for having high heterogeneity among the population.  
20 In addition, the fingerprint method may make use of the pattern of homologies indicated by a series of more and more stringent washes. Then, each position has both a sequence specificity and a homology measurement, the combination of which greatly increases the number of dimensions and the statistical  
25 likelihood of a perfect pattern match with another genetic individual.

### Genetic Screening

#### 1. test alleles with markers

30 The present invention provides for the ability to screen for genetic variations of individuals. For example, a number of genetic diseases are linked with specific alleles. See, e.g., Scriber, C. et al. (eds.) (1989) The Metabolic Bases of Inherited Disease, McGraw-Hill, New York. In one  
35 embodiment, cystic fibrosis has been correlated with a specific gene, see, Gregory et al. (1990) Nature 347: 382-386. A number of alleles are correlated with specific genetic deficiencies. See, e.g., McKusick, V. (1990) Genetic Inheritance in Man:

WO 92/10588

PCT/US91/09226

103

Catalogs of Autosomal Dominant, Autosomal Recessive, and X-linked Phenotypes, Johns Hopkins University Press, Baltimore; Ott, J. (1985) Analysis of Human Genetic Linkage, Johns Hopkins University Press, Baltimore; Track, R. et al. (1989) Banbury Report 32: DNA Technology and Forensic Science, Cold Spring Harbor Press, New York; each of which is hereby incorporated herein by reference.

## 2. Amniocentesis

Typically, amniocentesis is used to determine whether chromosome translocations have occurred. The mapping procedure may provide the means for determining whether these translocations have occurred, and for detecting particular alleles of various markers.

## MAPPING

### Positionally Located Clones

The present invention allows for the positional location of specific clones useful for mapping. For example, caged biotin may be used for specifically positioning a probe to a location on a matrix pattern.

In addition, the specific probes may be positionally directed to specific locations on a substrate by targeting. For example, polypeptide specific recognition reagents may be attached to oligonucleotide sequences which can be complementarily targeted, by hybridization, to specific locations on a VLSIPS substrate. Hybridization conditions, as applied for oligonucleotide probes, will be used to target the reagents to locations on a substrate having complementary oligonucleotides synthesized thereon. In another embodiment, oligonucleotide probes may be attached to specific polypeptide targeting reagents such as an antigen or antibody. These reagents can be directed towards a complementary antigen or antibody already attached to a VLSIPS substrate.

In another embodiment, an unnatural nucleotide which does not interfere with natural nucleotide complementary hybridization may be used to target oligonucleotides to

WO 92/10588

PCT/US91/09226.

104

particular positions on a substrate. Unnatural optical isomers of natural nucleotides should be ideal candidates.

In this way, short probes may be used to determine the mapping of long targets or long targets may be used to map the position of shorter probes. See, e.g., Craig et al. 1990  
5 Nuc. Acids Res. 18: 2653-2660.

#### Positionally Defined Clones

Positionally defined clones may be transferred to a new substrate by either physical transfer or by synthetic  
10 means. Synthetic means may involve either a production of the probe on the substrate using the VLSIPS synthetic methods, or may involve the attachment of a targeting sequence made by VLSIPS synthetic methods which will target that positionally  
15 defined clone to a position on a new substrate. Both methods will provide a substrate having a number of positionally defined probes useful in mapping.

#### CONCLUSION

The present inventions provide greatly improved methods and apparatus for synthesis of polymers on substrates. It is to be understood that the above description is intended to be illustrative and not restrictive. Many embodiments will be apparent to those of skill in the art upon reviewing the  
25 above description. By way of example, the invention has been described primarily with reference to the use of photoremovable protective groups, but it will be readily recognized by those of skill in the art that sources of radiation other than light could also be used. For example, in some embodiments it may be  
30 desirable to use protective groups which are sensitive to electron beam irradiation, x-ray irradiation, in combination with electron beam lithograph, or x-ray lithography techniques. Alternatively, the group could be removed by exposure to an electric current. The scope of the invention should,  
35 therefore, be determined not with reference to the above description, but should instead be determined with reference to the appended claims, along with the full scope of equivalents to which such claims are entitled.

WO 92/10588

PCT/US91/09226 .

105

All publications and patent applications referred to herein are incorporated by reference to the same extent as if each individual publication or patent application was specifically and individually incorporated by reference. The present invention now being fully described, it will be apparent to one of ordinary skill in the art that many changes and modifications can be made thereto without departing from the spirit or scope of the appended claims.

WO 92/10588

PCT/US91/09226

106

WHAT IS CLAIMED IS:

1. A composition comprising a plurality of positionally distinguishable sequence specific reagents attached to a solid substrate, which reagents are capable of specifically binding to a predetermined subunit sequence of a preselected multi-subunit length having at least five subunits, said reagents representing substantially all possible sequences of said preselected length.
2. A composition of Claim 1, wherein said subunit sequence is a polynucleotide.
3. A composition of Claim 1, wherein said specific reagent is an polynucleotide of at least about eight nucleotides.
4. A composition of Claim 1, wherein said specific reagents are all attached to a single solid substrate.
5. A composition of Claim 1, wherein said reagents comprise about 3000 different sequences.
6. A composition of Claim 1, wherein said reagents represents at least about 25% of the possible subsequences of said preselected length.
7. A composition of Claim 1, wherein said reagents are localized in regions of the substrate having a density of at least 25 regions per square centimeter.

WO 92/10588

PCT/US91/09226

107

8. A composition of Claim 4, wherein said substrate has a surface area of less than about 4 square centimeters.

9. A method of analyzing a sequence of a polynucleotide, said method comprising the step of:

a) exposing said polynucleotide to a composition of Claim 1.

10. A method of identifying or comparing a target sequence with a reference, said method comprising the step of:

a) exposing said target sequence to a composition of Claim 1;  
b) determining the pattern of positions of said reagents which specifically interact with said target sequence; and  
c) comparing said pattern with the pattern exhibited by said reference when exposed to said composition.

11. A method for sequencing a segment of a polynucleotide comprising the steps of:

a) combining:  
i) a substrate comprising a plurality of chemically synthesized and positionally distinguishable oligonucleotides capable of recognizing defined oligonucleotide sequences; and

WO 92/10588

PCT/US91/09226 .

108

ii) a target polynucleotide; thereby  
forming high fidelity matched duplex  
structures of complementary  
subsequences of known sequence; and

5           b) determining which of said reagents have  
specifically interacted with subsequences  
in said target polynucleotide.

12. A method of Claim 11, wherein said segment is  
10 substantially the entire length of said polynucleotide.

13. A method for sequencing a polymer, said method  
comprising the steps of:

15           a) preparing a plurality of reagents which  
each specifically bind to a subsequence of  
preselected length;

          b) positionally attaching each of said  
reagents to one or more solid phase  
substrates, thereby producing substrates of  
20 positionally definable sequence specific  
probes;

          c) combining said substrates with a target  
polymer whose sequence is to be determined;  
and

25           d) determining which of said reagents have  
specifically interacted with subsequences  
in said target polymer.

WO 92/10588

PCT/US91/09226

109

14. A method of Claim 13, wherein said substrates are beads.

15. A method of Claim 13, wherein said plurality of reagents comprise substantially all possible subsequences of said preselected length found in said target.

16. A method of Claim 13, wherein said solid phase substrates are a single substrate having attached thereto reagents recognizing substantially all possible subsequences of preselected length found in said target.

17. A method of Claim 13, further comprising the step of analyzing a plurality of said recognized subsequences to assemble a sequence of said target polymer.

18. A method of Claim 14, wherein at least some of said plurality of substrates have one subsequence specific reagent attached thereto, and said substrates are coded to indicate the specificity of said reagent.

19. A method of using a fluorescent nucleotide to detect interactions with oligonucleotide probes of known sequence, said method comprising:

- a) attaching said nucleotide to a target unknown polynucleotide sequence, and
- b) exposing said target polynucleotide sequence to a collection of positionally defined oligonucleotide probes of known

WO 92/10588

PCT/US91/09226

110

sequences to determine the sequences of  
said probes which interact with said  
target.

5           20. A method of Claim 19, further comprising the  
step of:

          a) collating said known sequences to determine  
              the overlaps of said known sequences to  
              determine the sequence of said target  
10               sequence.

          21. A method of mapping a plurality of sequences  
relative to one another, said method comprising:

          a) preparing a substrate having a plurality of  
15               positionally attached sequence specific  
              probes are attached;  
          b) exposing each of said sequences to said  
              substrate, thereby determining the patterns  
              of interaction between said sequence  
20               specific probes and said sequences; and  
          c) determining the relative locations of said  
              sequence specific probe interactions on  
              said sequences to determine the overlaps  
              and order of said sequences.

25

          22. A method of Claim 21, wherein said sequence  
specific probes are oligonucleotides.

WO 92/10588

PCT/US91/09226

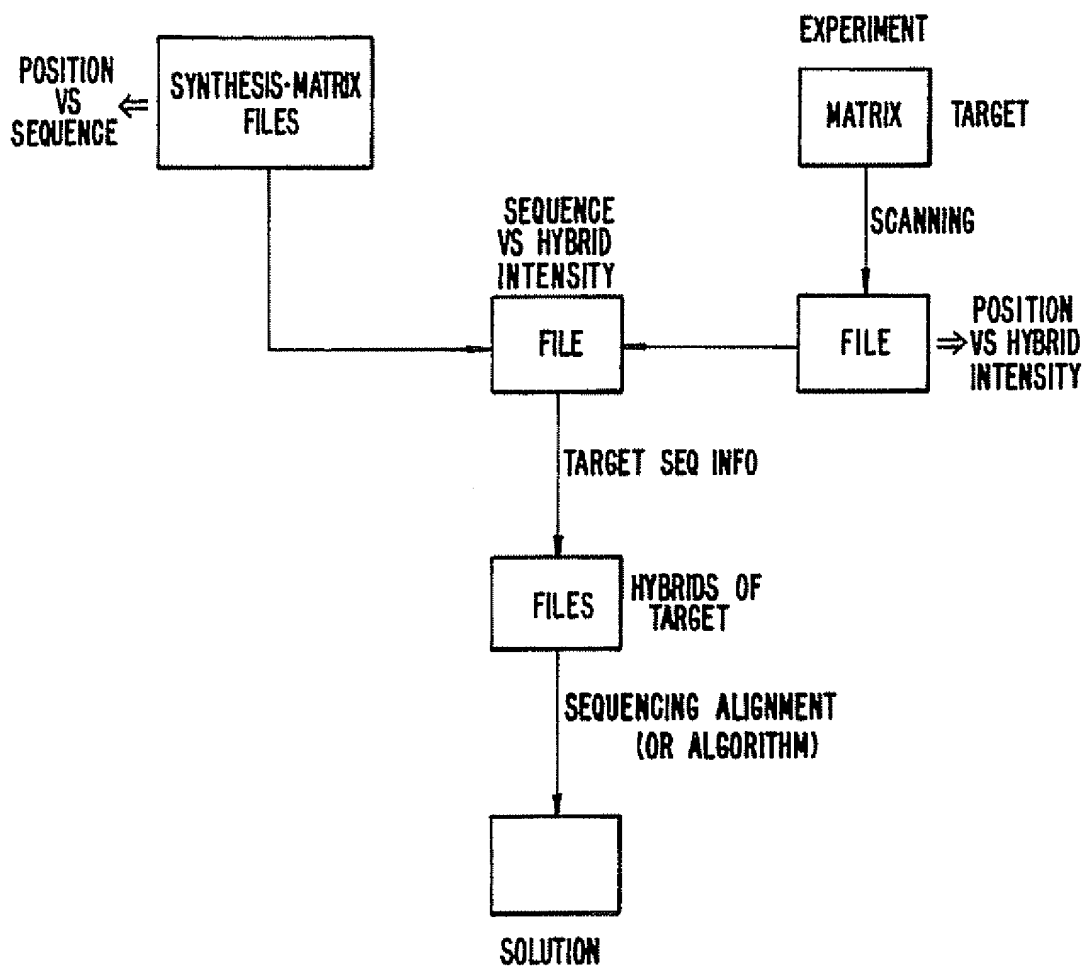
111

23. A method of Claim 21, wherein said sequences are nucleic acid sequences.

WO 92/10588

PCT/US91/09226

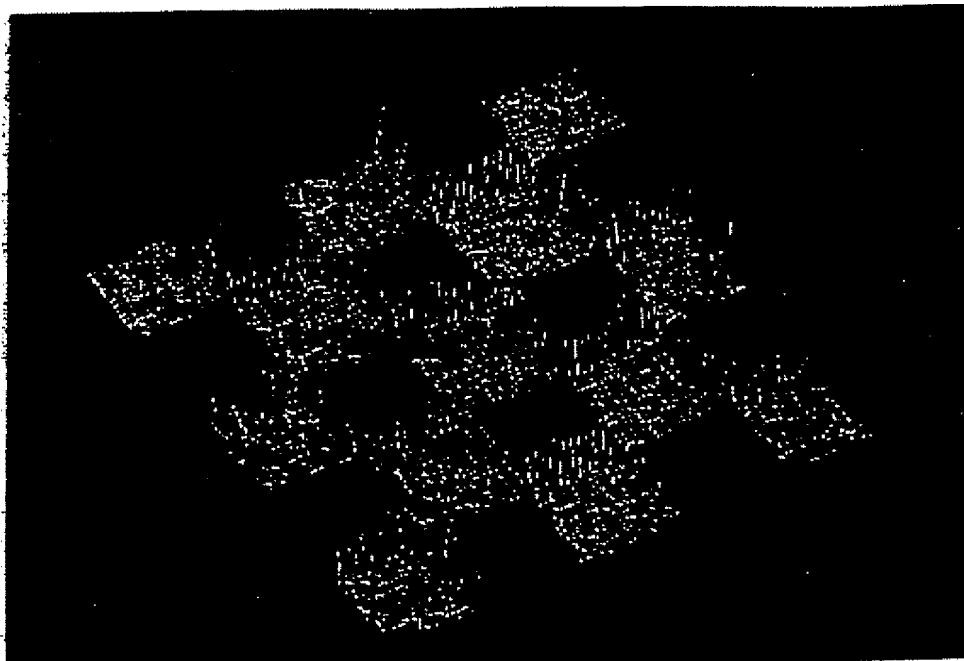
1/3

**FIG. 1.**

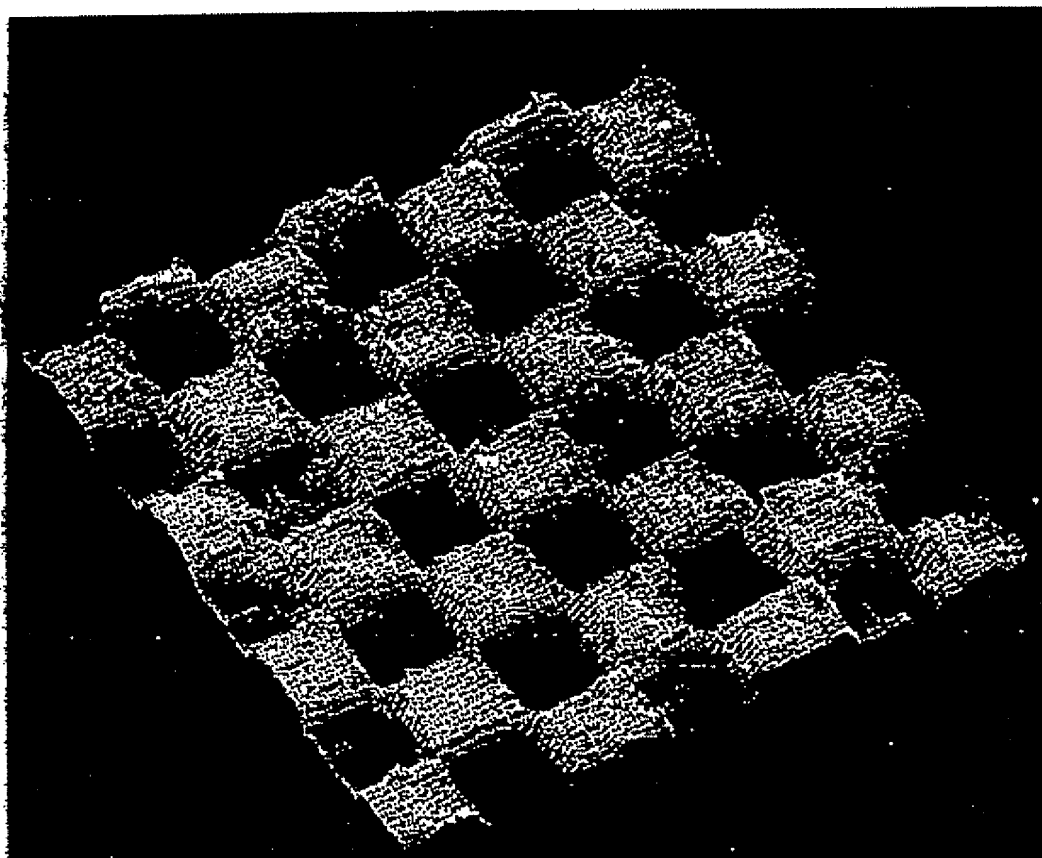
WO 92/10588

PCT/US91/09226

2/3



*FIG. 2.*



*FIG. 3.*

SUBSTITUTE SHEET

WO 92/10588

PCT/US91/09226

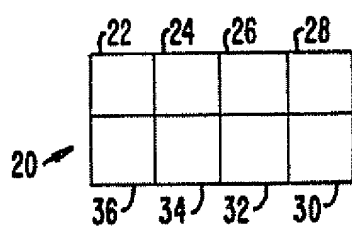


FIG. 4A.

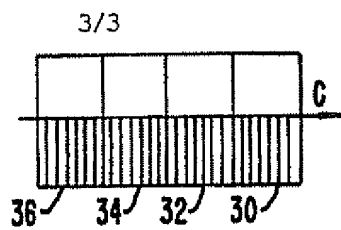


FIG. 4B.

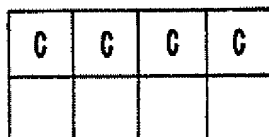


FIG. 4C.

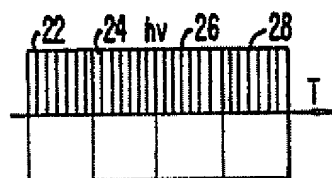


FIG. 4D.

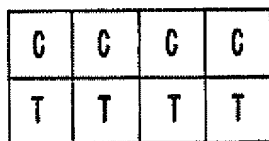


FIG. 4E.

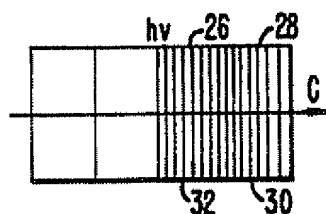


FIG. 4F.

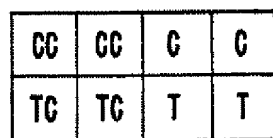


FIG. 4G.

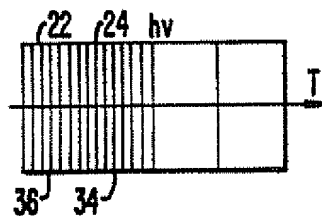


FIG. 4H.

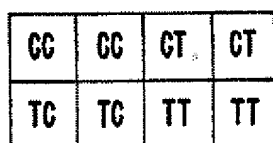


FIG. 4I.

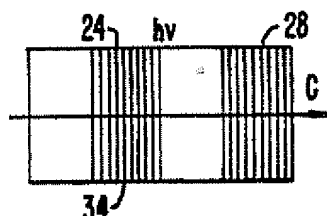


FIG. 4J.

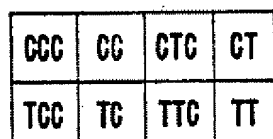


FIG. 4K.

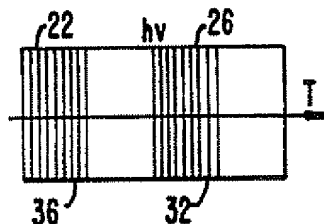


FIG. 4L.

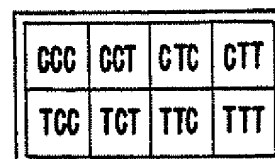


FIG. 4M.

## INTERNATIONAL SEARCH REPORT

International Application No. PCT/US91/09226

<b>I. CLASSIFICATION OF SUBJECT MATTER</b> (If several classification symbols apply, indicate all) <sup>3</sup>		
According to International Patent Classification (IPC) or to both National Classification and IPC		
IPC (5): C12Q 1/68; G01N 33/566, 33/48; C07H 15/12 US CL : 435/6; 436/501, 94; 536/26, 27, 28, 29; 935/77, 78		
<b>II. FIELDS SEARCHED</b>		
Minimum Documentation Searched <sup>4</sup>		
Classification System	Classification Symbols	
U.S.	435/6; 436/501, 94; 536/26, 27, 28, 29; 935/77, 78	
Documentation Searched other than Minimum Documentation to the extent that such Documents are included in the Fields Searched <sup>5</sup>		
Please See Attached Sheet.		
<b>III. DOCUMENTS CONSIDERED TO BE RELEVANT</b> <sup>14</sup>		
Category*	Citation of Document, <sup>16</sup> with indication, where appropriate, of the relevant passages <sup>17</sup>	Relevant to Claim No. <sup>18</sup>
X/Y	Doklady Akademii Nauk SSSR, Vol. 303, No. 6, issued December 1988, Yu. P. Lysov et al., "A New Method for Determining the DNA Nucleotide Sequence by Hybridization with Oligonucleotides", pages 436-438, see page 437, paragraph 3 and page 436, paragraph 2 (translation, Plenum Publishing Corporation, 1989).	1-13, 15-17, 18-23/14, 18
<p>* Special categories of cited documents:<sup>18</sup></p> <p>"A" document defining the general state of the art which is not considered to be of particular relevance</p> <p>"E" earlier document but published on or after the international filing date</p> <p>"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)</p> <p>"O" document referring to an oral disclosure, use, exhibition or other means</p> <p>"P" document published prior to the international filing date but later than the priority date claimed</p> <p>"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention</p> <p>"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step</p> <p>"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art</p> <p>"&amp;" document member of the same patent family</p>		
<b>IV. CERTIFICATION</b>		
Date of the Actual Completion of the International Search <sup>2</sup>	Date of Mailing of this International Search Report <sup>7</sup>	
10 MARCH 1992	17 MAR 1992	
International Searching Authority <sup>1</sup>	Signature of Authorized Officer <sup>20</sup>	
ISA/US	Stephanie W. Zitomer, Ph.D.	